WP 5.6: Support Vector Machine Imputation Methodology

H. Mallinson and A. Gammerman

Department of Computer Science
Royal Holloway, University of London
Egham, Surrey TW20 0EX
{alex,hugh}@cs.rhul.ac.uk

Contents

1 Introduction

This report is concerned with approaches to the 'missing data problem', specifically those known as imputation methods. The estimation of statistics from a dataset that is missing values is a common problem in National Statistics, the domain from which we draw our experimental datasets. We investigate a technique using Support Vector Machines (SVMs) for dealing with missing data in such surveys.

The technical details of the algorithm are described in a separate document. We will use this report to place the technique in the framework of existing approaches.

Imputation techniques solve the missing data problem by completing the dataset with 'plausible' surrogate values. We note that statistics can be estimated without imputation. This requires formulae to be handcrafted

for each dataset, a complex task requiring considerable expertise. Imputation has the advantage that once the dummy values have been inserted all standard, complete-data techniques can be applied.

Imputations (the inserted values) should ideally preserve the full joint probability of the data set. We use both measures of the preservation of marginal distributions and the preservation of true values to compare the methods on real world data sets. These measures of success compare the imputed values with the missing true values. The formulae for the performance measures are given in the third section.

The second section gives a formal definition of the maximum-likelihood model for datasets missing values. This is useful for clarifying the assumptions implicit in the various imputation approaches, and the patterns of missingness each can effectively treat, even those that do not attempt a model of the joint probability. It provides the conceptual framework for the EM and Data Augmentation Methods described in sections 5 and 6, which are based on parametric models of the full joint probability.

In the fourth section some simpler approaches to the missing data problem are described. These methods include list-wise deletion and mean imputation. The EM algorithm is presented in the fifth section. This iterative algorithm is used to estimate parametric models for the data. It is sensitive to the missing data pattern, and produces a maximum-likelihood solution. EM is related to the Bayesian method described in section 6, which also uses parametric models.

The Bayesian approach to imputation is also known as 'multiple imputation'. This is because the Bayesian approach allows one to make $k > 1$ draws $Y \sim P(Y_{miss}|Y_{obs})$ for each missing datum. Using a general set of formulae these multiple completions can be combined in any statistical estimate and importantly, measures of the variance of these statistics will naturally embody the extra uncertainty due to the incompleteness of the data set.

In section 7, an overview of Felligi-Holt techniques is given. These methods were developed for census data. Typically such data contains variables that must satisfy 'edit rules'. These capture logical relations existing between variables, for example, 'age' and 'marital status'.

Our aim is to assess Support Vector Machines which are state of the art prediction tools introduced in section 8. They able to learn complex dependencies between a set of input variables and a categorical or scalar output variable. They are straightforward to regularise, and make efficient use of the available data. As SVMs supply a conditional mean or mode, a separate predictor must be estimated for each missing variable. It is our

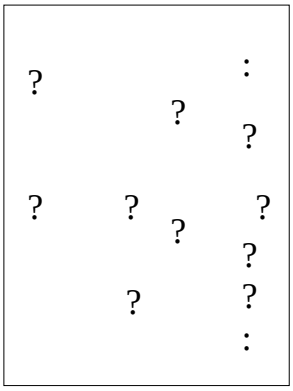

Figure 1: Incomplete data

supposition that non-linear correlations between subsets of the variable can be exploited by these algorithms. 'True values' will have been preserved only if a dependency or correlation between the variable with non-response and observed variables can be well estimated. However if no such dependencies exist we will expect the SVM to produce a mean imputation.

2 Imputation Problem

2.1 The data

The problem is represented graphically in figure 1. Here a dataset $A_{n \times m}$ of individuals (rows) measured on a number of variables (columns) is shown. The question marks represent values that are missing. If A were the data collected from a census, each row would represent a person, and each item on that row, the answer to a question put to that person. The rows, a_i are assumed to be exchangeable.

In its most general form the task is to enable users of the dataset to easily and accurately extract statistics, and measures of their precision, from the dataset. The job of imputation is to complete the dataset. I.e. to draw from a model of $P(Y_{miss}|Y_{obs})$, so that the statistics of the joint probability are preserved. We wish to preserve means, variances and correlations.

It may also be a goal of the imputation process to allow subsequent analysis to be sensitive to the missing information. *Multiple imputation*

techniques attempt to achieve this last aim.

2.2 Missing Data Patterns

The values missing are assumed to be produced by various types of 'missing data pattern'. If a variable is missing values completely at random (MCAR), a variable's absence is independent of all values observed or unobserved in the data set, including its own value. Missing at random (MAR) contains MCAR as a special case. A data set has a MAR missing data pattern if that pattern is independent of the missing values, given the observed values. In other words, if $P(R|Y_{obs}, Y_{miss}) = P(R|Y_{obs})$. Where R_{ij} represents a missingness indicator matrix. Not-missing-at-random (NMAR) describes cases where MAR does not hold. For example, if, in a survey rich people were unwilling to divulge information about their income, the pattern of non-reponse would clearly be dependent upon the values that were missing. Such patterns of missingness are known as non-ignorable. The observed values do not contain the information required to produce valid statistics for the full population.

In this work we evaluate algorithms on MAR and MCAR patterns. MCAR implies that any training set that can be assembled will be iid with the test set. MAR patterns are such that some subset of the data can be extracted such that the conditional distribution of any variables missing data given observed data will be iid in observed and missing data. MAR and MCAR patterns are collectively known as ignorable missingness mechanisms.

3 Evaluation Criteria

3.1 Scalar Variables

We measure the preservation of marginal distribution (the first two criteria) and the preservation of true values (the second two criteria). The second group contains measures used in standard machine learning scenarios; root mean square error and mean absolute error. The first type is concerned with measuring the distance of the whole set of true values from the whole set of imputations - *preservation of marginal distribution* compares the distributions of the set of imputations with the set of true values. In this case is not important whether each *individual* imputation is close to its true value. The performance is good if the set of imputations as a whole follows the

same distribution as the true values they replace. If the imputations were a random permutation of the true values the measure would give a perfect score, even if individually each imputation deviated strongly from the value it replaced.

The measures of preservation of true values,(mean absolute error and root mean square error) give an expected deviation of each imputed value from its true value. The preservation of true values is 'harder' and success will imply preservation of marginal distributions.

3.1.1 Preservation of marginal distribution

We compute the weighted empirical distribution functions for both sets of values. Y_i^* represents a true value, $\hat{Y}_i$ an imputation, w_i represent weights, all set to 1 here, $I()$ is an indicator function and n is the number of imputations made on one variable.

$$F_{Y^*n}(t) = \frac{\sum_{i=1}^{n} w_i I(Y_i^* \leq t)}{\sum_{i=1}^{n} w_i}$$

$$F_{\hat{Y}n}(t) = \frac{\sum_{i=1}^{n} w_i I(\hat{Y}_i \leq t)}{\sum_{i=1}^{n} w_i}$$

We then measure the 'distance' between these functions, in the two following ways.

1. Kolmogorov-Smirnov Distance **sMargKS**

   $$d_{KS}(F_{Y^*n}, F_{\hat{Y}n}) = max_t\left(|F_{Y^*n}(t) - F_{\hat{Y}n}(t)|\right) = max_j\left(|F_{Y^*n}(t_j) - F_{\hat{Y}n}(t_j)|\right)$$

   where the t_j are the $2n$ jointly ordered true and imputed values of Y.

2. Alpha distance metric **sMargAlph** :-

   $$d_\alpha(F_{Y^*n}, F_{\hat{Y}n}) = \frac{1}{t_{2n} - t_0} \sum_{j=1}^{2n} (t_j - t_{j-1}) |F_{Y^*n}(t_j) - F_{\hat{Y}n}(t_j)|^\alpha$$

   where α is a 'suitable' positive constant and t_0 is the largest integer smaller than or equal to t_1

   Normally $\alpha = 0$ or 1.

The measures both take values in the interval [0, 1, zero being best and one worst. The Kolmogorov Smirnov distance measures the maximum distance between the empirical cumulative distribution functions. The worst score occurs if a t value exists that scores zero on one of the functions and one on another. This happens if all imputations are smaller than all true values, or vice versa, i.e. if the cumulative distribution functions do not overlap at all.

3.1.2 Preservation of true values

We use two measures for preservation of true values for a scalar variable, The first is the mean absolute deviation, $stvd1$ and the second is the root mean square error, $stvd2$. The second is more sensitive to the presence of a few large errors. Y_i^* represents a true value, $\hat{Y}_i$ an imputation and n is the number of imputations.

$$stvd1 = \frac{\sum_{i=1}^{n} |(\hat{Y}_i^* - Y_i)|}{n}$$

$$stvd2 = \sqrt{\frac{\sum_{i=1}^{n} (\hat{Y}_i^* - Y_i)^2}{n}}$$

Both of these measure take values from 0 upwards. A mean imputation will have an error equal to the estimated variance of the observed values of the variable.

4 Standard techniques

4.1 Deletion of Units

This simple approach removes all rows lacking values on any variables. This is simple to implement and quick to execute. If few variables are missing, and the pattern is MCAR, it is reasonable. However if several variables lack values independently, the proportion of the dataset that is discarded can be high. For example, if four variables have a 5 % missingness, list-deletion would expect to remove nearly 20% $100 - (0.95^4) \times 100\%$) this would be a significant proportion of the data. In some datasets deletion is not an option, in a census for example.

4.2 Imputating overall mean or mode

Each unit with a missing item is imputed with the mean of the respondents, (for numeric variables) or the mode (for categorical variables). It is best used where only a small proportion of the data for each variable is missing. It embodies the assumption that variables are not correlated. The mean is retained, but the varaiance of the variable is reduced and any correlations with other variables can be dampened.

4.3 Group Mean

Each missing item is replaced by a group mean or mode. The groups are defined by fully observed discrete (or discretised) covariates. The covariates could be hand picked using apriori knowledge or 'learned' by an evolutionary technique for example. The technique is similar to donor imputation.

4.4 Hot Deck or Donor

Each missing item is replaced by a copying a randomly selected value from a subgroup ('donor pool') defined by one or more fully observed covariates. For example, on a census data set, one could impute income by copying from a randomly chosen respondent of the same sex and age.

4.5 Discussion of Standard Techniques

The techniques described in section 1 are not necessarily ad-hoc. Given that the data has been generated according to the right distribution they may the simplest to implement and introduce no distortion to the distributions. However they are normally applied without investigation into the assumptions they embody, or with any principled attempt to choose the covariates that define the best donor pool or group.

It should be noted that some techniques are problematic if unfeasible regions exist, (usually characterised by edit rules). Methods that attempt to generalise from the data, feature and group means for example, could impute unfeasible values.

In addition group mean and donor techniques are susceptible to the 'Curse of Dimensionality' [5]. 'Lattice-based' technique which model the data in local regions in the input space scale badly with input dimension.

5 EM Algorithm

Given observed data Y_{obs} and unobserved data Y_{miss}, the Expectation Maximisation (EM) algorithm finds θ that maximises $P(Y_{obs}|\theta)$. The EM algorithm iteratively produces a sequence of parameter estimates $\theta_1, \theta_2, \ldots$ that converge to a local maximum of the observed data likelihood. It is assumed that the observed data likelihood is diffiuclt to handle directly, but the joint distribution $P(Y_{obs}, Y_{miss}|\theta)$ can be maximised analytically. After a random initialisation of the model θ the two following steps are repeated,

- E step: Estimate $\hat{P} = P(Y_{miss}|Y_{obs}, \theta_{old})$
- M step: $\theta_{new} = max_\theta[E_{\hat{P}}[P(Y_{obs}, Y_{miss}|\theta)]]$

In the first step a distribution over the missing data is calculated, given a prior estimate of the model. This is then used in the second step to reestimate the model. The approach is appropriate for any MCAR and MAR missing data pattern, in a dataset that is of the parametric form of θ. EM has been used to estimate the the parameters of mixture models. There is no missing data in this situation, but extra parameters, θ_2(mixture components) can be introduced which take on the role of the missing data in the algorithm. This approach is taken because the likelihood $P(data|\theta)$ is intractable. The stationary points in the derivative are difficult to find. $P(data, \theta_2|\theta_1)$ is much easier to handle. The method provably finds a local maxima of the likelihood function. In many cases it is also a global maxima. Given that the data conforms to the parametric model, the method will work for ignorable missing data patterns. The method can be adapted to work with mixture models for missing data problems also. Convergence is easier to diagnose than for the full Bayesian methods described in the next chapter. Because the method only prodcues a point estimate, confidence limits are not available for the imputations. The dependence on the parametric form can be limiting if their are non-linear correlations in the data. The process could get trapped in a local minima. The standard reference is [13], as this was the first point at which convergence was proved . Examples of the algorithm had been seen much earlier however. Recent extensions by Tresp and Gharamani use EM with mixture models and incomplete data, and multi-layer perceptrons. Neal and Hinton have investigated generalisations of the algorithm in which data and parameters are updated component at a time. This has been found to speed convergence.

6 Bayesian Methods for Imputation

6.1 Introduction

Multiple Imputation (MI) results from a Bayesian approach to modelling the dataset. MI responds to the need for methods that capture the additional uncertainty due to missingness. It is a simulation based technique in which each missing dataum is replaced with a set of $m > 1$ values. The m versions of the dataset are analyzed with standard complete data methods. The m results are then combined with simple rules and produce estimates of standard errors and p-values that formally incorporate missing data uncertainty.

6.2 Simulation of probability density functions

Data Augmentation developed by Tanner and Wong is a *simulation technique* used in the application of full Bayesian methods to the missing data problem. Bayesian methods can require the integration of intractable probability density functions. Various methods have been developed to approximate such integrals, for example variational or Markov chain Monte Carlo techniques. The latter family of methods use simulated draws from the intractable density function. Data Augmentation is an MCMC techniqe closely related to the more well known Gibbs Sampling method.

Ideally we would like to produce

$$P(Y_{mis}|Y_{obs}) = \int P(Y_{mis}|\theta)P(\theta|Y_{obs})d\theta$$

and draw values of Y_{mis} to complete the data set. However as suggested above, the posterior distribution of θ cannot be handled analytically. The integral is approximated by simulating the second distribution on the right hand side. Simulation involves producing a large number number of draws from the intractable distribution. I.e. a number of θ's are produced $\theta_1, \theta_2, ...\theta_3, ...\theta_n$ according to the posterior distribution. The integral can then be approximated:

$$P(Y_{mis}|Y_{obs}) \approx \frac{1}{n}\sum_i P(Y_{mis}|\theta_i)$$

6.3 Gibbs Sampling

We will sketch Gibb's sampling as it offers a clearer presentation of the ideas than the closely related technique of data augmentation. Gibb's sampling

exploits the situation in which our multidimensional target random variable cannot be simulated, but conditional distributions of its subvectors can. Let random vector $Z = (Z_1, Z_2, ..Z_J)$ have joint distribution $P(Z)$ as our target to be simulated. We iteratively draw from the conditional distribution of each subvector given all the others.

$$Z_1^{t+1} \sim P(Z_1|Z_2^t, ...Z_J^t)$$

$$Z_2^{t+1} \sim P(Z_2|Z_1^{t+1}, Z_3^t ...Z_J^t)$$

$$Z_3^{t+1} \sim P(Z_3|Z_1^{t+1}, Z_2^{t+1} ...Z_J^t)$$

After we have done t complete cycles through the J conditional vectors we have a sequence

$$Z^1, Z^2, ...Z^t (= (Z_1^t, Z_2^t, ...Z_J^t)$$

that forms a Markov chain, which under certain conditions [1] has a stationary distribution equla to $P(Z)$.

6.4 Strengths and Weaknesses

Multiple Imputation offers the only principled method for incorporating the uncertainty due to missing data. It is dependent however on a number of assumptions concerning the data, the prior distribution of the model parameters and mechanism of non-response.

The first assumption is that the the data conforms to the parametric model chosen. The second concerns a problem inherent to Bayesian approaches. The effect of the prior over the model parameters may be quite strong if the dataset is not large, or if certain variables have high missingness. Lastly, if the mechanism is not MAR the MI will, like all standard methods fail.

7 NIM and Felligi-Holt

The nearest-neighbour imputation model (NIM) has been developed specifically for the census setting. This setting assumes not all errors have been *localised*. Variables that are in a strict logical relationship can be checked

[1] The regularity conditions necessary to establish convergence of the Gibbs sampler are technical. They do tend to be satisfied in most problems of practical interest

for consistency. In a household data set for example, the respondents that have age less than sixteen must also have 'single' as their marital status. If a unit fails to be consistent, (it fails an 'edit rule'), the task is to first pick which variable to impute, marital status or age and then to impute a legal value. This is called the *edit and imputation* scenario.

It should be noted that the data sets typically tackled in this scenario contain $O(10^6)$ records. Household data is, in addition, normally large, noisy, multimodal and O(100) dimensionality.

Felligi-Holt approaches, which pre-date NIM, the edit rules are processed to identify the *minimum-change-set*. This is the smallest set of variables that could be altered to render the unit consistent. Donor methods are subsequently used to find a 'plausible' value.

In NIM, the changes the algorithm makes depend upon which donors are available to impute from. If respondent A is 14 years old and divorced, a donor set of variables is first assembled, by identifying a number of 'nearest neighbours' under one change set are much closer than than under the others this is identified as the optimal change set. Under Felligi-Holt no account is made of the likelihood of the imputed unit. By using information from the donor pool one can approximate a 'most likely' change-set, as opposed to a 'minimum-change' set. These techniques can handle large data sets and strict edit rules, localising errors as well as imputing legal values. The techniques are suited to scenarios in which the data is well understood, and neighbourhood metrics are easy to define.

8 Imputation with Support Vector Machines

8.1 Introduction

A separate document contains a description of the SVM imputation harness, and the core SVM algorithms. Here we motivate an SVM approach to imputation and place it in the context of existing techniques. In particular we shall argue that many of the issues to be considered when using SVMs are common to feed-forward multi-layer perceptrons.

The SVM approach proposed treats imputation as a generalisation of *prediction*. The goal in prediction is estimate target values such that a unit level measure of error is minimisied, for example the root-mean square error. This is not the only measure of error that is relevant in missing data problems however. We discuss the implications of treating imputation as prediction, and particulars of our 'meta-approach' in the first section.

In the second section we review the issues that prompted the development of the SVM. We present features of the algorithm that recommend their application in standard classification and regression problems. We investigate whether these features can be exploited in the more general missing-data scenario also.

The approach proposed by Rubin, using models of the full joint probability will be known as the 'model-based' approach. Imputation methods using models of the conditional expectation will be known collectively as 'prediction approaches'. Prediction approaches include some neural networks models, the group-mean algorithm, hot-deck[2] and SVMs. We will distinguish issues that we believe to be common to all prediction approaches, and those that pertain only to the SVM.

8.2 Imputation with Support Vector Machines

8.2.1 Predicting Missing Values

Imputation should preserve the full joint distribution $P(X)$, of the data. Devising measures for the preservation of P is difficult. Hence, various components of the full distribution are used as proxies. For example, the Kolmogorov-Smirnoff distance is used to measure the preservation of the marginal distribution of each missing variable, and root-mean-square error is used to measure the preservation of true values. Using these measures we can compare the performance of various approaches. Techniques that perform well on one measure however, might not be so good at others.

Our SVM imputation harness extracts a number of prediction problems from the data set, treating them sequentially. An SVM predicts a univariate target value from a (fully observed) multivariate training vector, assuming i.i.d data, producing a model for the conditional expectation; $E(X_{target}|X_1, ..., X_n)$. A parameterisation is found that minimises the root-mean-square-error on the training set, subject to regularisation term.

Hence, by handling imputation with a prediction algorithm we effectively chose preservation of true values as the primary goal. The implicit assumption is that the conditional distribution is unimodal and conditional variance is low, and hence that most data lies near their expected value, conditioned on the other variables.

Imputing with the conditional expectation will tend to reinforce any

[2]hot-deck attempts to maintain the variance in the distribution by drawing imputations from a pool of data 'near' to the conditional expectation

correlations in the data and compresses the distribution of the target variable. Neural nets can be considered as prediction algorithms and will also accentuate correlations.

In the next section we discuss the generalisation of the SVM to several missing variables. The points made in this section concerning prediction methods and their affect on the joint probability distribution will be developed in the conclusions section. We proceed under the assumption that imputing with the conditional expectation is acceptable.

8.2.2 Generalising SVM to Several Missing Variables

We can immediately apply SVM to a non-hierarchical data set, that has only one variable missing values[3]. I.e. missing-data problems reduce to standard regression or classification problems if only one variable is missing.

The situation is more complicated when many variables are missing values[4]. Whichever variable is first imputed, a strategy for dealing with missing *input* variables must be found. We have investigated two such approaches to this problem.

heuristic1 extracts a fully observed subset of the data for training. Missing input variables on the test units are estimated by a mean or modal value.

heuristc2 estimates all missing covariates with a mean or mode. 'Patched' data is used for training.

The first heuristic may not be possible at all as there may be no units that are fully observed. It will usually result in a much reduced training set. If two variables are independently MCAR missing at a 20% rate, the expected proportion of complete units is 64%. If 5 variables are missing in this way there will be just 33% of the data available.

Heuristic1 also assumes that the missing data pattern for the target variable is MCAR (missing completely at random[22]). For example, consider a hypothetical survey scenario in which students always leave some questions unanswered. Training a model to predict *income* using only fully observed units will not access information from students. It is likely that imputations for *income* for students will be inaccurate, as the model will not be well determined for students.

[3] e.g. Danish Labour Force Survey, only *income* variable missing values

[4] e.g. Sample of Anonymised Records (SARS) data set. Many of the variables are missing values, and a single unit may lack as many as five simultaneously

The second heuristic presently uses crude methods to estimate missing data in the training and test set input variables. This may distort any relationships that do exist in the data.

8.3 Motivation for Support Vector Machines

Support Vector Machines, introduced by Boser, Guyon and Vapnik[2], are just one technique in a group known as 'Kernel Methods'. These algorithms merge concepts from statistics, functional analysis, optimisation and machine learning. They are often non-linear generalisations of pre-existing linear techniques, exploiting an implicit projection of the data to a high-dimensional feature space, provided by the kernel function.

In addition to the SVM regression and classification tools, a kernel PCA [3] has been studied and a novelty detection algorithm [9]. Campbell developed a kernelised Fisher's discriminant algorithm, and Saunders et al devised a kernelised ridge regression [11].

The appeal of Kernel Methods lies firstly in their access to a rich set of non-linear models. Like neural networks they are in principle universal approximators. This means that, given enough data the algorithm can produce a regressor or discriminant to fit any continuous surface. Secondly, kernel methods offer algorithmic efficiency. Training[5] requires the solution of a convex quadratic program. Such problems have been extensively studied and efficient methods exist for solving them. Testing is linear in the size of test set.

Thirdly, due to the simplicity of the underlying linear algorithm, attractive theoretical properties can be proved, bounds on the generalisation error for example. We discuss this aspect in the third section.

The good performance of SVMs, particularly on high-dimensional problems, is believed to be based firstly in their avoidance of density estimation and secondly, in their efficient parameterisation. We explore these issues further in the next section.

In the last section, we discuss two successful applications of SVM classification; hand-written digit recognition and text retrieval. By contrasting these problem domains with the problem in hand, imputation in National Statistics, we hope to draw some qualitative conclusions.

[5] otherwise known as parameter estimation

8.3.1 Density estimation and the Curse of Dimensionality

Vapnik [27] argued that 'when solving a given problem, try to avoid solving a more general problem as an intermediate step'. Density estimation is more general than classification and regression in the following sense; if we know the full joint p.d.f. $P(X,Y)$, we can derive $P(Y|X)$. However density estimation is an example of an *ill-posed* problem: given a small deviation in the sample, large deviations in the estimated parameters of the model may result.

Vapnik argued that regression and classification problems should be solved *directly*, without estimation of probability densities. SVMs implement this philosophy.

In addition, SVMs can be seen to handle some of the problems that befall classical parametric and non-parametric approaches when applied in high-dimensions. These problems are known collectively as 'The Curse of Dimensionality'[5].

One problem exemplified by multivariate polynomial regression is known as *parameter explosion*. The number of parameters for an Mth order model grows like d^M, a power law dependence on the dimension d. Large training sets are required in order that the model be well determined[7].

Models that perform local density estimates are also problematic in high dimensions. The nearest-neighbours and Parzen window models are examples in point. If we place a hypercubical box around a point in a 10 dimensional space, containing 1% of the data (assumed to be uniformly distributed), the expected edge length will be $(0.01)^{\frac{1}{10}} = 0.63$. The box will hence have side length of over 60% of the range of each input variable. The attempt to provide a local model fails.

The SVM avoids density estimation. Moreover, through use of the kernel functions, a rich non-linear hypothesis space is supplied with no increase in the number of parameters. Regularisation, through maximisation of the margin, is also independent of the dimension.

8.3.2 Bounding the Generalisation Error

Consider a classification problem. The goal is to find $f : \mathbb{R}^N \to \{\pm 1\}$ using input-output training data that is independently and identically distributed,

$$(\mathbf{x}_1, y_1), ..., (\mathbf{x}_l, y_l) \in \mathbb{R}^N \times \{\pm 1\}$$

such that f will correctly classify unseen examples $(\mathbf{x}_{l+1}, y_{l+1})$ i.e. $f(\mathbf{x}_{l+1}) = y_{l+1}$, where the test pair is drawn from the same distribution as the training

set.

The training error of a given classifier is given by,

$$err_{train}[f] = \frac{1}{2\,l}\sum_{i=1}^{l} |f(\mathbf{x}_i) - y_i|$$

and the true error[6] is given by,

$$err_{true}[f] = \frac{1}{2}\int |f(\mathbf{x}) - y|\, dP(\mathbf{x}, y)$$

The true error cannot be calculated as we do not know the distribution, $P(\mathbf{x}, y)$. VC theory devises methods for calculating the probability p of the true error deviating by more than a given amount from the training error;

$$\delta = P\{|err_{true} - err_{train}| > \epsilon\}$$

We can interpret δ as the probability of being *mislead* by the training performance. Investigation into how δ could be bounded lead to the discovery of *capacity concepts* such as the VC-Dimension[7]. Capacity concepts give a measure of the flexibility of a family of models. A family with high capacity will have a member that can fit any training set closely. The 1-nearest-neighbours model, for example, will always give zero training error. However, if there is noise we are likely to overfit. Such high capacity models therefore have large δ.

The technique of regularisation in classical statistics is closely related to capacity. However capacity measures produce an integer value for a family that can be related, with the training error and size of data set, to a bound on δ, for a given ϵ. More usefully, the bound is rearranged. ϵ is given in terms of a given δ, h the VC dimension of the family of models and l the size of the training set,

$$\epsilon \leq \sqrt{\frac{h(log\frac{2l}{h} + 1) - log(\delta/4)}{l}}$$

If δ were set to 0.05, the bound states that there is only a 1 in 20 chance that the true error is more than ϵ away from the training error.

The margin in an SVM acts as a capacity measure. It should be maximised to minimise capacity and give tighter bounds. Of two functions achieving the same training error, the better one has the larger margin since the true error is more likely to be close to the training error.

[6] also known as the risk functional

[7] defined as the largest number h, of points that can be separated in all possible ways using functions of the given class.

8.3.3 Applications of Kernel Methods

SVMs have performed impressively on hand-written digit recognition problems. The publicly available USPS data set contains 10,000 handwritten numerals, described as grey-scale vectors in a $20\times\ 20 = 400$ pixel space. The goal is to automate the process of sorting mail by Zip code[28].

The SVM was able to give near optimal performance on USPS *without* knowledge of the geometry of the problem. Existing techniques with comparable performance exploit a priori knowledge.

This application domain is characterised by its high dimensionality, and low noise. Classes of handwritten digits overlap only to a small degree. It should also be noted that these sorts of problems do not require a classifier that is transparent, i.e. that can justify or explain its decisions. All that is required is an accurate classification.

SVMs have also performed well on text categorisation[17] problems where the goal is to sort documents into a number of predefined categories according to content. This problem is characterised by an input space with thousands of dimensions, and low noise. The documents are vectorised by considering each word in the language as a dimension.

The domain of National Statistics presents datasets of relatively low dimensions, with many categorical variables. Many of the variables are related logically, for example marital status and age. Common sense dictates that the most subtle dependencies in the variables are still unlikely to require highly non-linear models.

This considerations must limit our expectations of SVMs outperforming more transparent models. We will take these issues up again in the conclusions.

9 Summary

9.1 SVM Imputation

The support vector machine offers non-linear, univariate prediction. It has been harnessed to the imputation problem in an ad-hoc manner, estimating the missing values by training a set of models, one SVM for each variable missing values. The problem thus becomes one of how to best select training data for each model. The order in which variables are imputed is also important, as imputations for one variable may subsequently be used as training data for other missing data.

Model selection is carried out through cross-validation. For the rbf kernel[8] 3 parameters must be chosen before seeing the data. Good settings can be found by considering about 120 different combinations, taking 2 hours approximately on our platform.

9.2 Results on DLFS

Results on the Danish Labour Force Survey data set (DLFS),[9] showed the SVM improving upon the performance of the MLP and the group-mean algorithms. We qualified our conclusions however, noting that the benchmarks were not tuned as extensively as the SVM. For example we investigated only one approach to partitioning scalar variables. Results also indicated that the SVM may be mislead by noisy redundant variables.

9.3 Results on SARS

Initial experiments with the Sample of Anonymised Records (SARS) indicate that the SVM performs similarly to the group-mean and other algorithms. On some multiclass problems it offered better preservation of distributions. However a fuller investigation of feature extraction and pre-processing is necessary before confident conclusions can be made.

9.4 Conclusions

We make the following tentative remarks regarding the nature of the data that is being imputed and the suitability of SVMs. These remarks are based on the results observed on two datasets, and certain background knowledge concerning the items (people, businesses) represented by the data.

9.4.1 Simple Dependencies

The datasets in the Euredit Project are large, and low dimensional. There is enough data to estimate conditional probabilities locally. Many variables are categorical, for example *sex* and *marital status*, with few values. Particular variables are known apriori to bear (simple) relationships to each other, for example marital status and age. In some cases these relationships are clearly deductive. In other words, an appreciation of the entities and the variables

[8] The choice of rbf kernel is not necessarily optimal. Other kernels should also be investigated

[9] Presented in a separate document

they are measured on shows the data to contain simple dependencies that fit well with apriori common sense.

In summary, we do not expect complex unkown multivariate dependencies to exist in these data sets, that can *only* be captured by algorithms such as the SVM or the MLP, fitting highly non-linear dependencies. Stratification and some simple transformations should usually be adequate.

9.4.2 Expected Gains in Performance

Hence we believe SVMs are unlikely to offer a large improvement over simpler approaches, which are more transparent and quicker to apply. Moreover, in some situations linear and group mean methods appear to be more robust to noisy, redundant variables.

9.4.3 Apriori knowledge and Edit Rules not handled

Moreover the SVM has not been implemented in a way that can integrate knowledge of structure in the data set. It is possible to impute rules that contradict edit rules for example. Hierarchical structure, for example in household surveys, can also not be exploited at present.

9.4.4 Usability

The SVM is presently a piece of research software, and accepts data only in a flat, csv format. Preprocessing is required. We investigate normalisation and the use of design variables. It can offer only the training error, and validation results in general, as diagnostic information. The model is non-probabilistic, giving point predictions without confidence or credibility ratings. The model does not indicate which features are most important for the prediction task.

References

[1] V. Vovk A. Gammerman and V. N. Vapnik. Learning by transduction. In G. Cooper and S. Moral, editors, *Uncertainty in Artificial Intelligence Proceedings of the 14th Conference*, pages 148–155, 1998.

[2] I. M. Guyon B. E. Boser and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152, 1992.

[3] A. Smola B. Schölkopf and K. Müller. Kernel principal component analysis. *Advances in Kernel Methods - Support Vector Learning*, 1999.

[4] Bankier. Nearest neighbour imputation methodology. Technical report, Canadian Office of Statistics, 1998.

[5] R. E. Bellman. *Adaptive Control Processes*. Princeton UNiversity Press, 1961.

[6] Y. Bengio and F. Gingras. Recurrent neural networks for missing or asynchronous data. *Advances in Neural Information Processing Systems*, 1996.

[7] C. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.

[8] G. E. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. J. Wiley and Sons, New York, wiley classics library edition edition, 1992.

[9] B. Schölkopf C. Burges and V. Vapnik. Extracting the support data for a given task. *Proceedings First International Conference on Knowledge Discovery and Data Mining*, 1995.

[10] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Knowlege Discovery and Data Mining*, pages 100 – 130, 1998.

[11] A. Gammerman C. Saunders and V. Vovk. Ridge regression learning algorithm in dual variables. *Machine Learning: Proceeding of the Fifteenth International Conference*, 1999.

[12] N. Cristianini and Shawe-Taylor J. *An Introduction to Support Vector Machines*. CUP, 2000.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society B*, 39:1 – 38, 1977.

[14] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall, 1997.

[15] Zoubin Ghahramani and Michael I. Jordan. Learning from incomplete data. Technical Report AIM-1509, 1994.

[16] R. Herbrich, T. Graepel, and C. Campbell. Bayes point machines: Estimating the bayes point in kernel space, 1999.

[17] T. Joachims. Text categorization with support vector machines. In *Proceedings of European Conference on Machine Learning (ECML)*, 1998.

[18] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data.* Wiley, New York, 1987.

[19] B. A. Murtagh and M. A. Saunders. Minos 5.1 user's guide. Technical Report SOL-83-20R, Stanford University, CA, USA, 1983.

[20] J. Platt. Fast training of support vector machines using sequential minimal optimzation. *Advances in Kernel Methods - Support Vector Learning*, 1996.

[21] D. B. Rubin. http://www.statsol.ie. SOLAS 2.0.

[22] D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys.* Wiley Series in Probability and Mathematical Statistics. Wiley, New York, 1987. ISSN: 0271-6232.

[23] J. L. Schafer. *Analysis of Incomplete Multivariate Data.* Number 72 in Monographs on Statistics and Applied Probability. Chapman and Hall, London, 1997. ISBN: 0412040611.

[24] Alex J. Smola and Bernhard Schölkpf. A tutorial on support vector regression. Technical Report NC2-TR-1998-030, GMD, 1998.

[25] Volker Tresp, Subutai Ahmad, and Ralph Neuneier. Training neural networks with deficient data. In Jack D. Cowan, Gerald Tesauro, and Joshua Alspector, editors, *Advances in Neural Information Processing Systems*, volume 6, pages 128–135. Morgan Kaufmann Publishers, Inc., 1994.

[26] R. J. Vanderbei. Loqo: An interior point code for quadratic programming. Technical Report SOR-94-15, Statistics and Perations Research, Princeton University, 1994.

[27] V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, 1995.

[28] L. Bottou Y. LeCun, L. D. Jackel and A. Brunot. Comparison of learning algorithms for handwritten digit recognition. In F. Fogelman-Soulie and P. Gallinari, editors, *Proceedings ICANN'95 - International Conference on Artificial Neural Networks*, pages 53–60, 1995.