Learning from Incomplete Data with Infinite Imputations

Uwe Dick DICK@MPI-SB.MPG.DE
Peter Haider HAIDER@MPI-SB.MPG.DE
Tobias Scheffer SCHEFFER@MPI-SB.MPG.DE
Max Planck Institute for Computer Science, Saarbrücken, Germany

Abstract

We address the problem of learning decision functions from training data in which some attribute values are unobserved. This problem can arise, for instance, when training data is aggregated from multiple sources, and some sources record only a subset of attributes. We derive a generic joint optimization problem in which the distribution governing the missing values is a free parameter. We show that the optimal solution concentrates the density mass on finitely many imputations, and provide a corresponding algorithm for learning from incomplete data. We report on empirical results on benchmark data, and on the email spam application that motivates our work.

1. Introduction

In many applications, one has to deal with training data with incompletely observed attributes. For instance, training data may be aggregated from different sources. If not all sources are capable of providing the same set of input attributes, the combined training sample contains incompletely observed data. This situation occurs in email spam detection, where it is helpful to augment the content of an email with real-time information about the sending server, such as its blacklist status. This information is available for all training emails that arrive at a mail server under one's own control, and it is also available at application time. But if one wants to utilize training emails from public archives, this information is missing.

We adress a learning setting in which values are *missing at random:* here, the presence or absence of values does not convey information about the class labels. If this condition is not met, it is informative to consider the presence or absence of values as additional input to the decision function. Techniques for learning from incomplete data typically involve a distributional model that imputes missing values, and the desired final predictive model. Prior work on learning from incomplete data is manifold in the literature, and may be grouped by the way the distributional model is used.

The first group models the distribution of missing values in a first step, and learns the decision function based on the distributional model in a second step. Shivaswamy et al. (2006) formulate a loss function that takes a fixed proportion of the probability mass of each instance into account, with respect to the estimated distribution of missing values. They derive second order cone programs which renders the method applicable only to very small problems. Other examples include Williams and Carin (2005), Williams et al. (2005), and Smola et al. (2005).

The second group estimates the parameters of a distributional model and the final predictive model jointly. As an example, recently Liao et al. (2007) propose an EM-algorithm for jointly estimating the imputation model and a logistic regression classifier with linear kernel, assuming the data arises from a mixture of multivariate Gaussians.

The third group makes no model assumption about the missing values, but learns the decision function based on the visible input alone. For example, Chechik et al. (2007) derive a geometrically motivated approach. For each example, the margin is re-scaled according to the visible attributes. This procedure specifically aims at learning from data with values that are *structurally missing*—as opposed to *missing at random*. Chechik et al. (2007) find empirically that the procedure is not adequate when values are missing at random.

Jointly learning a distributional model and a kernel predictive model relates to the problem of learning a

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

kernel function from a prescribed set of parameterized kernels. This problen drew a lot of attention recently; see, for example, Argyriou et al. (2005) and Micchelli and Pontil (2007).

Estimating the distributional model first and training the predictive model in a second step leaves the user free to choose any learning algorithm for this second step. However, a harder problem has to be solved than would be necessary. If one is only interested in a decision function that minimizes the desired loss, knowing the values or distribution of the missing attributes in the training set is not actually required. Furthermore, errors made in the imputation step and errors made in estimating the parameters of the predictive model can add up in a sequential procedure.

Consequently, we investigate learning the decision function and the distribution of imputations dependently. Unlike prior work on this topic, we develop a solution for a very general class of optimization criteria. Our solution covers a wide range of loss functions for classification and regression problems. It comes with all the usual benefits of kernel methods. We derive an optimization problem in which the distribution governing the missing values is a free parameter. The optimization problem searches for a decision function and a distribution governing the missing values which together minimize a regularized empirical risk.

No fixed parametric form of the distributional model is assumed. A regularizer that can be motivated by a distributional assumption may *bias* the distributional model *towards* a prior belief. However, the regularizer may be overruled by the data, and the resulting distributional model may be different from any parametric form. We are able to prove that there exists an optimal solution based on a distribution that is supported by finitely many imputations. This justifies a greedy algorithm for finding a solution. We derive manifestations of the general learning method and study them empirically.

The paper is structured as follows. After introducing the problem setting in Section 2, we derive an optimization problem in Section 3. Section 4 proves that there is an optimal solution that concentrates the density mass on finitely many imputations and presents an algorithm. Example instantiations of the general solution are presented in Section 5. We empirically evaluate the method in Section 6. Section 7 concludes.

2. Problem Setting

We address the problem of learning a decision function f from a training sample in which some attribute values are unobserved.

Let $\mathbf{X}$ be a matrix of n training instances $\mathbf{x}_i$ and let $\mathbf{y}$ be the vector of corresponding target values y_i. Instances and target values are drawn *iid* from an unknown distribution $p(\mathbf{x}, y)$ with $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathcal{Y}$, where $\mathcal{Y}$ denotes the set of possible target values. Matrix $\mathbf{Z}$ indicates which features are observed. A value of $z_{il} = 1$ indicates that x_{il}, the l-th feature of the i-th example, is observed. Values are *missing at random:* y_i is conditionally independent of $\mathbf{z}_i$ given $\mathbf{x}_i$.

The goal is to learn a function $f : \mathbf{x} \mapsto y$ that predicts target values for *completely observed* examples. The decision function should incur only a minimal true risk $R(f) = \int L(y, f(\mathbf{x}))p(\mathbf{x}, y)d\mathbf{x}dy$, where L is a loss function for the task at hand.

As a means to minimizing the true risk, we seek a function f in the *reproducing kernel Hilbert space* $\mathcal{H}_k$ induced by a kernel k that minimizes a regularized empirical risk functional $R(f) = \sum_{i=1}^{n} l(y_i, f(\mathbf{x}_i)) + \eta\|f\|_k^2$. We demand k to be a Mercer kernel. Loss function l approximates the true loss L. The *representer theorem* allows us to write the minimizer as a sum over functions in $\mathcal{H}_k$ centered at training instances: $f(\mathbf{x}) = \sum_{j=1}^{n} c_j k(\mathbf{x}_j, \mathbf{x})$.

The learning problem from completely observed data would amount to solving Optimization Problem 1.

Optimization Problem 1 (Primal learning problem, observed data). *Over* $\mathbf{c}$*, minimize*

$$R(\mathbf{c}, k) = \sum_{i=1}^{n} l\Big(y_i, \sum_{j=1}^{n} c_j k(\mathbf{x}_j, \mathbf{x}_i)\Big) + \eta \sum_{i,j=1}^{n} c_i c_j k(\mathbf{x}_j, \mathbf{x}_i)$$

We require that the loss function be defined in such a way that Optimization Problem 1 can be written in the dual form of Optimization Problem 2. A wide range of loss functions satisfies this demand; we will later see that this includes hinge loss and squared loss.

Optimization Problem 2 (Dual of learning problem). *Given* $a < 0$*, over* $\mathbf{c}$*, maximize*

$$a \langle \mathbf{c}, \mathbf{K}\mathbf{c} \rangle - R^*(\mathbf{c})$$

subject to the constraints

$$\forall_{i=1}^{m_1^*} g_i^*(\mathbf{c}) \leq 0, \qquad \forall_{j=1}^{m_2^*} h_j^*(\mathbf{c}) = 0. \tag{1}$$

$R^*(\mathbf{c})$ denotes a differentiable convex function of the dual variables $\mathbf{c}$ which we demand to be independent of the kernel matrix $\mathbf{K}$. The inequality constraints g_i^* are differentiable convex and the equality constraints h_j^* differentiable affine. We like to note that the requirement of independence between R^* and $\mathbf{K}$ is not

very restrictive in practice, as we will see in chapter 5. Furthermore, we demand strong duality to hold between Optimization problems 1 and 2.

3. Learning from Incomplete Data in One Step

If any instance $\mathbf{x}_i$ has unobserved features, then $k(\mathbf{x}_i, \mathbf{x})$ and, consequently, the decision function f are not properly defined. In order to learn from incomplete data, we will marginalize the decision function and risk functional by the observable attributes and integrate over all unobserved quantities. To this end, we define $\boldsymbol{\omega} \in \Omega_{\mathbf{X}}^{\mathbf{Z}} \subset \mathbb{R}^{n \times d}$ as a matrix of imputations constrained by $\omega_{il} = x_{il}$ if $z_{il} = 1$. We demand $\Omega_{\mathbf{X}}^{\mathbf{Z}}$ to be compact for the rest of this paper. Let $\boldsymbol{\omega}_i$ denote the i-th row of $\boldsymbol{\omega}$. Then we can define a family of kernels $K(\boldsymbol{\omega})(\mathbf{x}_j, \mathbf{x}_i) = k(\boldsymbol{\omega}_j, \boldsymbol{\omega}_i)$. Any probability measure $p(\boldsymbol{\omega})$ on imputations induces a marginalization of the kernel by the observable variables. Equation 2 integrates over all imputations of unobserved values; it can be evaluated based on the observed values.

$$K(p)(\mathbf{x}_j, \mathbf{x}_i) = \int_{\boldsymbol{\omega} \in \Omega_{\mathbf{X}}^{\mathbf{Z}}} k(\boldsymbol{\omega}_j, \boldsymbol{\omega}_i) dp(\boldsymbol{\omega}) \tag{2}$$

Any probability measure $p(\boldsymbol{\omega})$ constitutes an optimization criterion $R(\mathbf{c}, K(p))$. In the absence of knowledge about the true distribution of missing values, $p(\boldsymbol{\omega})$ becomes a free parameter. Note that $p(\boldsymbol{\omega})$ is a continuous probability measure that is not constrained to any particular parametric form; the space of parameters is therefore of infinite dimensionality.

It is natural to add a regularizer $Q(p)$ that reflects prior belief on the distribution of imputations $p(\boldsymbol{\omega})$ to the optimization criterion, in addition to the empirical risk and regularizer on the predictive model. The regularizer is assumed to be continuous in p. The regularizer does not *constrain* $p(\boldsymbol{\omega})$ to any specific class of distribution, but it reflects that some distributions are believed to be more likely. Without a regularizer, the criterion can often be minimized by imputations which move instances with missing values far away from the separator, thereby removing their influence on the outcome of the learning process. This leads to Optimization Problem 3.

Optimization Problem 3 (Learning problem with infinite imputations). *Given n training examples with incomplete feature values, $\gamma > 0$, kernel function k, over all $\mathbf{c}$ and p, minimize*

$$\tilde{R}_{k,\gamma}(\mathbf{c}, p) \quad = \quad R(\mathbf{c}, K(p)) + \gamma Q(p) \tag{3}$$

subject to the constraints

$$\forall \boldsymbol{\omega} : \; p(\boldsymbol{\omega}) \geq 0, \qquad \int_{\boldsymbol{\omega} \in \Omega_{\mathbf{X}}^{\mathbf{Z}}} p(\boldsymbol{\omega}) d\boldsymbol{\omega} = 1.$$

Each solution to Optimization Problem 3 integrates over *infinitely* many different imputations. The search space contains all continuous probability measures on imputations, the search is guided by the regularizer Q. The regularization parameter γ determines the influence of the regularization on the resulting distribution. For $\gamma \to \infty$ the solution of the optimization reduces to the solution obtained by first estimating the distribution of missing attribute values that minimizes the regularizer. For $\gamma \to 0$ the solution is constituted by the distribution minimizing the risk functional R.

4. Solving the Optimization Problem

In this section, we devise a method for efficiently finding a solution to Optimization Problem 3. Firstly, we show that there exists an optimal solution $\hat{\mathbf{c}}, \hat{p}$ with $\hat{p}$ supported on at most $n+2$ imputations $\boldsymbol{\omega} \in \Omega_{\mathbf{X}}^{\mathbf{Z}}$. Secondly, we present an algorithm that iteratively finds the optimal imputations and parameters minimizing the regularized empirical risk.

4.1. Optimal Solution with Finite Combination

In addition to the parameters $\mathbf{c}$ of the predictive models, continuous probability measure $p(\boldsymbol{\omega})$ contributes an infinite set of parameters to Optimization Problem 3. The implementation of imputations as parameters of a kernel family allows us to show that there exists an optimal probability measure $\hat{p}$ for Equation 3 such that $\hat{p}$ consists of finitely many different imputations.

Theorem 1. *Optimization Problem 3 has an optimal solution $\hat{\mathbf{c}}, \hat{p}$ in which $\hat{p}$ is supported by at most $n+2$ imputations $\boldsymbol{\omega} \in \Omega_{\mathbf{X}}^{\mathbf{Z}}$.*

Proof. The compactness of $\Omega_{\mathbf{X}}^{\mathbf{Z}}$ and the continuity of $\mathbf{K}$ immediately imply that *there exists some* solution to Optimization Problem 3. It remains to be shown that at least one of the solutions is supported by at most $n+2$ imputations. Let $\bar{\mathbf{c}}, \bar{p}$ be *any* solution and let all requirements of the previous section hold. The idea of this proof is to construct a correspondence between distributions over imputations and vectors in $\mathbb{R}^{n+1}$, where a finite support set is known to exist. Define $\mathbf{S}(\boldsymbol{\omega}) = \mathbf{K}(\boldsymbol{\omega})\bar{c} \in \mathbb{R}^n$ and $D = \{(\mathbf{S}(\boldsymbol{\omega})^\top, Q(\boldsymbol{\omega}))^\top : \boldsymbol{\omega} \in \Omega_{\mathbf{X}}^{\mathbf{Z}}\} \subset \mathbb{R}^{n+1}$. Since $\Omega_{\mathbf{X}}^{\mathbf{Z}}$ is compact and $K(\cdot)$ and $Q(\cdot)$ are continuous by definition, D is compact as well. We define a measure over D as $\mu(A \times B) = \bar{p}(\{\boldsymbol{\omega} : \mathbf{S}(\boldsymbol{\omega}) \in A \wedge Q(\boldsymbol{\omega}) \in B\})$.

Then, by Carathéodory's convex hull theorem, there exists a set of k vectors $\{(\mathbf{s}_1^\top, q_1)^\top, \ldots, (\mathbf{s}_k^\top, q_k)^\top\} \subseteq D$ with $k \leq n+2$ and nonnegative constants ν_i with

$\sum_{i=1}^{k} \nu_i = 1$, such that

$$\int_D (\mathbf{s}^\top, q)^\top d\mu((\mathbf{s}^\top, q)^\top) = \sum_{i=1}^{k} (\mathbf{s}_i^\top, q_i)^\top \nu_i.$$

For each i, select any $\boldsymbol{\omega}_i$ such that $(\mathbf{S}(\boldsymbol{\omega}_i)^\top, Q(\boldsymbol{\omega}_i)) = (\mathbf{s}_i^\top, q_i)$. We construct $\hat{p}$ by setting $\hat{p}(\boldsymbol{\omega}) = \sum_{i=1}^{k} \nu_i \delta_{\boldsymbol{\omega}_i}$, where $\delta_{\boldsymbol{\omega}_i}$ denotes the Dirac measure at $\boldsymbol{\omega}_i$. The optimal $\hat{\mathbf{c}}$ results as $\arg\min_{\mathbf{c}} R(\mathbf{c}, K(\hat{p}))$. We have

$$\begin{aligned} \int_D \mathbf{s} d\mu((\mathbf{s}^\top, q)^\top) &= \sum_{i=1}^{k} \mathbf{s}_i \nu_i, \quad \text{and} \\ \int_D q d\mu((\mathbf{s}^\top, q)^\top) &= \sum_{i=1}^{k} q_i \nu_i. \end{aligned}$$

Then

$$\begin{aligned} \mathbf{K}(\bar{p})\bar{\mathbf{c}} &= \left(\int_{\Omega_{\mathbf{X}}^{\mathbf{Z}}} \mathbf{K}(\boldsymbol{\omega}) d\bar{p}(\boldsymbol{\omega}) \right) \bar{\mathbf{c}} = \int_{\Omega_{\mathbf{X}}^{\mathbf{Z}}} \mathbf{S}(\boldsymbol{\omega}) d\bar{p}(\boldsymbol{\omega}) \\ &= \int_D \mathbf{S}(\boldsymbol{\omega}) d\mu \left((\mathbf{S}(\boldsymbol{\omega})^\top, Q(\boldsymbol{\omega}))^\top \right) \\ &= \sum_{i=1}^{k} \mathbf{s}_i \nu_i = \int_{\Omega_{\mathbf{X}}^{\mathbf{Z}}} \mathbf{S}(\boldsymbol{\omega}) d\hat{p}(\boldsymbol{\omega}) \\ &= \int_{\Omega_{\mathbf{X}}^{\mathbf{Z}}} \mathbf{K}(\boldsymbol{\omega}) d\hat{p}(\boldsymbol{\omega}) \bar{\mathbf{c}} = \mathbf{K}(\hat{p})\bar{\mathbf{c}}. \end{aligned}$$

Likewise,

$$\begin{aligned} Q(\bar{p}) &= \int_D Q(\boldsymbol{\omega}) \, d\mu \left((\mathbf{S}(\boldsymbol{\omega})^\top, Q(\boldsymbol{\omega}))^\top \right) \\ &= \sum_{i=1}^{k} q_i \nu_i = Q(\hat{p}). \end{aligned}$$

Since $Q(p)$ does not depend on $\mathbf{c}$, $\bar{\mathbf{c}} = \arg\min_{\mathbf{c}} R(\mathbf{c}, \mathbf{K}(\bar{p}))$, and by strong duality, $\bar{\mathbf{c}} = \arg\max_{\mathbf{c}} a \langle \mathbf{c}, \mathbf{K}(\bar{p})\mathbf{c} \rangle - R^*(\mathbf{c})$. This implies that the Karush-Kuhn-Tucker conditions hold for $\bar{\mathbf{c}}$, namely there exist constants $\kappa_i \geq 0$ and λ_j such that

$$a\mathbf{K}(\bar{p})\bar{\mathbf{c}} - \nabla R^*(\bar{\mathbf{c}}) + \sum_i \kappa_i \nabla g_i^*(\bar{\mathbf{c}}) + \sum_j \lambda_j \nabla h_j^*(\bar{\mathbf{c}}) = 0$$

$$\forall_i \;\; g_i^*(\bar{\mathbf{c}}) \leq 0, \quad \forall_j \;\; h_i^*(\bar{\mathbf{c}}) = 0, \quad \forall_i \;\; \kappa_i g_i^*(\bar{\mathbf{c}}) = 0$$

It is easy to see that therefore $\bar{\mathbf{c}}$ is also a maximizer of $a \langle \mathbf{c}, \mathbf{K}(\hat{p})\mathbf{c} \rangle - R^*(\mathbf{c})$, because $\mathbf{K}(\bar{p})\bar{\mathbf{c}} = \mathbf{K}(\hat{p})\bar{\mathbf{c}}$ and the Karush-Kuhn-Tucker conditions still hold. Their sufficiency follows from the fact that $\mathbf{K}(p)$ is positive semi-definite for any p, and the convexity and affinity premises. Thus,

$$\begin{aligned} & R(\bar{\mathbf{c}}, K(\bar{p})) + \gamma Q(\bar{p}) \\ &= \left[\min_{\mathbf{c}} R(\mathbf{c}, K(\bar{p})) \right] + \gamma Q(\bar{p}) \\ &= \left[\max_{\mathbf{c}} a \langle \mathbf{c}, \mathbf{K}(\bar{p})\mathbf{c} \rangle - R^*(\mathbf{c}) \right] + \gamma Q(\bar{p}) \\ &= [a \langle \bar{\mathbf{c}}, \mathbf{K}(\bar{p})\bar{\mathbf{c}} \rangle - R^*(\bar{\mathbf{c}})] + \gamma Q(\bar{p}) \\ &= [a \langle \bar{\mathbf{c}}, \mathbf{K}(\hat{p})\bar{\mathbf{c}} \rangle - R^*(\bar{\mathbf{c}})] + \gamma Q(\hat{p}) \\ &= \left[\max_{\mathbf{c}} a \langle \mathbf{c}, \mathbf{K}(\hat{p})\mathbf{c} \rangle - R^*(\mathbf{c}) \right] + \gamma Q(\hat{p}) \\ &= \left[\min_{\mathbf{c}} R(\mathbf{c}, K(\hat{p})) \right] + \gamma Q(\hat{p}) \\ &= R(\hat{\mathbf{c}}, K(\hat{p})) + \gamma Q(\hat{p}). \end{aligned}$$

We have now established that there exists a solution with at most $n + 2$ imputations. □

4.2. Iterative Optimization Algorithm

This result justifies the following greedy algorithm to find an optimal solution to Optimization Problem 3. The algorithm works by iteratively optimizing Problem 1 (or, equivalently, 2), and updating the distribution over the missing attribute values. Let $p_{\bar{\boldsymbol{\omega}}}$ denote the distribution $p(\boldsymbol{\omega}) = \delta_{\bar{\boldsymbol{\omega}}}$. Algorithm 1 shows the steps.

Algorithm 1 Compute optimal distribution of imputations on $\Omega_{\mathbf{X}}^{\mathbf{Z}}$

Initialization: Choose $p^{(1)} = p_{\boldsymbol{\omega}^{(1)}}$; e.g., $\boldsymbol{\omega}_{il}^{(1)} = 0$ for all $z_{il} \neq 1$
for $t = 1 \ldots$ **do**
1. $\hat{\mathbf{c}} \leftarrow \arg\min_{\mathbf{c}} R(\mathbf{c}, K(p^{(t)}))$
2. Find $\boldsymbol{\omega}^{(t+1)} \in \Omega_{\mathbf{X}}^{\mathbf{Z}}$: $\tilde{R}_{k,\gamma}(\hat{\mathbf{c}}, p_{\boldsymbol{\omega}^{(t+1)}}) < \tilde{R}_{k,\gamma}(\hat{\mathbf{c}}, p^{(t)})$. If no such $\boldsymbol{\omega}^{(t+1)}$ exists, terminate.
3. $\beta_t \leftarrow \arg\min_{\beta \in (0,1]} \left[\min_{\mathbf{c}} \tilde{R}_{k,\gamma}(\mathbf{c}, \beta p_{\boldsymbol{\omega}^{(t+1)}} + (1-\beta)p^{(t)}) \right]$
4. $p^{(t+1)} \leftarrow \beta_t p_{\boldsymbol{\omega}^{(t+1)}} + (1-\beta_t)p^{(t)}$
5. $\forall j < t : \; \beta_j \leftarrow \beta_j(1-\beta_t)$
end for

Step 1 consists of minimizing the regularized empirical risk functional R, given the current distribution. In step 2 a new imputation is constructed which improves on the current objective value. Since in general $\tilde{R}_{k,\gamma}(\mathbf{c}, p_{\boldsymbol{\omega}})$ is not convex in $\boldsymbol{\omega}$, one cannot find the *optimal* $\boldsymbol{\omega}$ efficiently. But the algorithm only requires to find *any* better $\boldsymbol{\omega}$. Thus it is reasonable to perform gradient ascent on $\boldsymbol{\omega}$, with random restarts in case the found local optimum does not satisfy the inequality of step 2. In step 3 and 4 the optimal distribution consisting of the weighted sum of currently used Dirac impulses $\sum_{i=1}^{t} \beta_i \delta_{\boldsymbol{\omega}_i}$ and the new imputation $\delta_{\boldsymbol{\omega}^{(t+1)}}$ is computed. This step is convex in β if

$\tilde{R}_{k,\gamma}(\mathbf{c}, \beta p_{\boldsymbol{\omega}^{(t+1)}} + (1-\beta)p^{(t)})$ is linear in β. By looking at Optimization Problem 2, we see that this is the case for R. Thus the convexity depends on the choice for Q (see Sect. 5.2). Step 5 updates the weights of the previous imputations.

The algorithm finds t imputations $\boldsymbol{\omega}^{(j)}$ and their weights β_j, as well as the optimal example coefficients $\mathbf{c}$. We can construct the classification function f as

$$f(\mathbf{x}) \quad = \quad \sum_{j=1}^{t} \sum_{i=1}^{n} \beta_j \mathbf{c}_i k(\boldsymbol{\omega}_i^{(j)}, \mathbf{x}). \tag{4}$$

Note that the value $n+2$ is an upper bound for the number of basic kernels which constitute the optimal solution. The algorithm is not guaranteed to terminate after $n+2$ iterations, because the calculated imputations are not necessarily optimal. In practice, however, the number of iterations is usually much lower. In our experiments, the objective value of the optimization problem converges in less than 50 iterations.

5. Example Learners

In this chapter we present manifestations of the generic method, which we call *weighted infinite imputations*, for learning from incomplete data that we use in the experimental evaluation.

Recall from Section 3 the goal to learn a decision function f from incomplete data that minimizes the expected risk $R(f) = \int L(y, f(\mathbf{x}))p(\mathbf{x}, y)d\mathbf{x}dy$. In classification problems the natural loss function L becomes the *zero-one loss*, whereas in regression problems the loss depends on the specific application; common choices are the *squared error* or the ϵ-*insensitive* loss. The considerations in the previous chapters show that, in order to learn regression or classification functions from training instances with missing attribute values, we only have to specify the dual formulation of the preferred learning algorithm on complete data and a regularizer on the distribution of imputations p.

5.1. Two Standard Learning Algorithms

For binary classification problems, we choose to approximate the zero-one by the *hinge* loss and perform *support vector machine* learning. The dual formulation of the SVM is given by $R^{SVM}(\mathbf{c}, k) = \sum_{i=1}^{n} \frac{c_i}{y_i} - \frac{1}{2}\sum_{i,j=1}^{n} c_i c_j k(\mathbf{x}_j, \mathbf{x}_i)$ subject to the constraints $0 \leq \frac{c_i}{y_i} \leq \frac{1}{\eta}$ and $\sum_{i=1}^{n} c_i = 0$. We see that the demands of Optimization Problem 2 are met and a finite solution can be found. Taking the SVM formulation as the dual Optimization Problem 2 gives us the means – in conjunction with an appropriate regularizer Q – to learn a classification function f from incomplete data.

For regression problems, the loss depends on the task at hand, as noted above. We focus on penalizing the *squared error*, though we like to mention that the approach works for other losses likewise. One widely used learning algorithm for solving the problem is *kernel ridge regression*. Again, we can learn the regression function f from incomplete data by using the same principles as described above. Kernel ridge regression minimizes the regularized empirical risk $\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2 + \eta\|f\|^2$. The dual formulation $R^{KRR}(\mathbf{c}, k) = \sum_{i=1}^{n} c_i y_i - \frac{1}{4}\sum_{i=1}^{n} c_i^2 + \frac{1}{4\eta}\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j)$ again meets the demands of the dual optimization problem 2. Substituting its primal formulation for R in step 1 of Algorithm 1 and in Eqn. 3 solves the problem of learning the regression function from incomplete data after specifying a regularizer Q.

5.2. Regularizing towards Prior Belief in Feature Space

A regularizer on the distribution of missing values can guide the search towards distributions $\hat{\boldsymbol{\omega}}$ that we believe to be likely. We introduce a regularization term which penalizes imputations that are different from our prior belief $\hat{\boldsymbol{\omega}}$. We choose to penalize the sum of squared distances between instances $\mathbf{x}_i$ and $\hat{\boldsymbol{\omega}}_i$ in *feature space* $\mathcal{H}_k$ induced by kernel k. We define the squared distance regularization term Q^{sq} as

$$\begin{aligned} Q^{sq}(k, \hat{\boldsymbol{\omega}}) \quad &= \quad \sum_{i=1}^{n} \|\phi_k(\mathbf{x}_i) - \phi_k(\hat{\boldsymbol{\omega}}_i)\|_2^2 \\ &= \quad \sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{x}_i) - 2k(\mathbf{x}_i, \hat{\boldsymbol{\omega}}_i) + k(\hat{\boldsymbol{\omega}}_i, \hat{\boldsymbol{\omega}}_i). \end{aligned}$$

Note that when using Q^{sq}, step 3 of Algorithm 1 becomes a convex minimization procedure.

5.3. Imputing the Mean in Feature Space

In principle any imputation we believe is useful for learning a good classifier can be used as $\hat{\boldsymbol{\omega}}$. Several models of the data can be assumed to compute corresponding optimal imputations. We like to mention one interesting model, namely the class-based mean imputation in the *feature space* $\mathcal{H}_k$ induced by kernel k. This model imputes missing values such that the sum of squared distances between completed instances to the class-dependent mean in feature space is minimal over all possible imputations. $\hat{\boldsymbol{\omega}} = \arg\min_{\boldsymbol{\omega}} \sum_{i=1}^{n} \|\phi_k(\boldsymbol{\omega}_i) - \frac{1}{n_{y_i}}\sum_{j:y_j=y_i} \phi_k(\boldsymbol{\omega}_j)\|_2^2$, where n_y denotes the number of instances with label y. Simple algebraic manipulations show that this is equivalent to

minimizing the sum of squared distances between all instances $\sum_{v\in\{-1,1\}} \frac{1}{n_v} \sum_{i,j:y_i=y_j=v} \|\phi_k(\boldsymbol{\omega}_i) - \phi_k(\boldsymbol{\omega}_j)\|_2^2 = \sum_{v\in\{-1,1\}} \frac{1}{n_v} \sum_{i,j:y_i=y_j=v} \big[k(\boldsymbol{\omega}_i,\boldsymbol{\omega}_i) - 2k(\boldsymbol{\omega}_i,\boldsymbol{\omega}_j) + k(\boldsymbol{\omega}_j,\boldsymbol{\omega}_j)\big]$

Definition 1 (Mean in Feature Space). *The class-based* mean in feature space *imputation method imputes missing values $\hat{\boldsymbol{\omega}}$ which optimize*

$$\hat{\boldsymbol{\omega}} = \arg\min_{\boldsymbol{\omega}} \sum_{v\in\{-1,+1\}} \frac{1}{n_v} \sum_{i,j:y_i=y_j=v} \big[k(\boldsymbol{\omega}_i,\boldsymbol{\omega}_i) - 2k(\boldsymbol{\omega}_i,\boldsymbol{\omega}_j) + k(\boldsymbol{\omega}_j,\boldsymbol{\omega}_j)\big]$$

Note that this model reduces to the standard mean in input space when using the linear kernel.

6. Empirical Evaluation

We evaluate the performance of our generic approach *weighted infinite imputations* for two example realizations. We test for classification performance on the email spam data set which motivates our investigation. Furthermore, we test on seven additional binary classification problems and three regression problems.

6.1. Classification

We choose to learn the decision function for the binary classification task by substituting the risk functional of the *support vector machine*, $-R^{SVM}$, as presented in section 5.1 for R and the squared distance regularizer Q^{sq} (Section 5.2) for Q in Optimization Problem 3.

For the motivating problem setting, we assemble a data set of 2509 spam and non-spam emails, which are preprocessed by a linear text classifier which is currently in use at a large webspace hosting company. This classifier discriminates reasonably well between spam and non-spam, but there is still a small fraction of misclassified emails. The classifier has been trained on about 1 million emails from a variety of sources, including spam-traps as well as emails from the hosting company itself, recognizing more than 10 million distinct text features. On this scale, training a support vector machine with Gaussian kernel is impractical, therefore we employ a two-step procedure. We discard the contents of the emails and retain only their spam score from the text classifier and their size in bytes as content features in the second-step classifier. At the time of collection of the emails, we record auxiliary real-time information about the sending servers. This includes the number of valid and invalid receiver addresses of all emails seen from the server so far, and the mean and standard deviation of the sizes and spam scores of all emails from the server. Such information is not available for emails from external sources, but will be available when classifying unseen emails. We randomly draw 1259 emails, both spam and non-spam, with server information, whereas half of those were drawn from a set of misclassified spam-emails. We augment this set with 1250 emails drawn randomly from a source without server information for which only 2 of the 8 attributes are observed.

To evaluate the common odd versus even digits discrimination, random subsets of 1000 training examples from the USPS handwritten digit recognition set are used. We test on the remaining 6291 examples. Additionally, we test on KDD Cup 2004 Physics (1000 train, 5179 test, 78 attributes) data set and on the 4-view land mine detection data (500, 213, 41) as used by Williams and Carin (2005). In the latter, instances consist of 4 views on the data, each from a separate sensor. Consequently, we randomly select complete views as missing. From the UCI machine learning repository we take the Breast (277 instances, 9 features), Diabetes (768, 8), German (1000, 20), and Waveform (5000, 21) data sets. Selection criteria for this subset of the repository were minimum requirements on sample size and number of attributes.

On each data set we test the performance of *weighted infinite imputation* using four different regularization imputations $\hat{\boldsymbol{\omega}}$ for the regularizer $Q^{sq}(K(p),\hat{\boldsymbol{\omega}})$. These imputations are computed by *mean imputation in input space* (**MeanInput**) and *mean imputation in feature space* (**MeanFeat**) as by Definition 1. Additionally we use the *EM* algorithm to compute the attributes imputed by the maximum likelihood parameters of an assumed multivariate Gaussian distribution with no restrictions on the covariate matrix (**Gauss**), and a Gaussian Mixture Model with 10 Gauss centers and spherical covariances (**GMM**).

Four learning procedures based on single imputations serve as reference methods: the **MeanInput**, **MeanFeat**, **Gauss**, and **GMM** reference methods first determine a single imputation, and then invoke the learning algorithm.

All experiments use a spheric Gaussian kernel. Its variance parameter σ as well as the SVM-parameter η are adjusted using the regular SVM with a training and test split on fully observed data. All experiments on the same data set use this resulting parameter setting. Results are averaged over 100 runs were in each run training and test split as well as missing attributes are chosen randomly. If not stated otherwise, 85% of attributes are marked missing on all data sets. In order to evaluate our method on the email data set, we perform 20-fold cross-validation. Since the emails with

Table 1. Classification accuracies and standard errors for all data sets. Higher accuracy values are written in bold face, "$*$" denotes significant classification improvement.

		MeanInput	Gauss	GMM	MeanFeat
Email	Single imp	0.9571 ± 0.0022	0.9412 ± 0.0037	0.9505 ± 0.0030	0.9570 ± 0.0022
	WII	0.9571 ± 0.0022	$\mathbf{0.9536 \pm 0.0022}$ $*$	$\mathbf{0.9527 \pm 0.0024}$	$\mathbf{0.9600 \pm 0.0019}$ $*$
USPS	Single imp	0.8581 ± 0.0027	0.8688 ± 0.0022	0.9063 ± 0.0012	0.8581 ± 0.0027
	WII	$\mathbf{0.8641 \pm 0.0027}$ $*$	$\mathbf{0.8824 \pm 0.0024}$ $*$	$\mathbf{0.9105 \pm 0.0015}$ $*$	$\mathbf{0.8687 \pm 0.0027}$ $*$
Physics	Single imp	0.6957 ± 0.0035	0.5575 ± 0.0038	0.6137 ± 0.0050	0.6935 ± 0.0028
	WII	$\mathbf{0.7084 \pm 0.0039}$ $*$	$\mathbf{0.6543 \pm 0.0055}$ $*$	$\mathbf{0.6881 \pm 0.0049}$ $*$	$\mathbf{0.7036 \pm 0.0032}$ $*$
Mine	Single imp	0.8650 ± 0.0025	0.8887 ± 0.0023	0.8916 ± 0.0023	0.8660 ± 0.0026
	WII	$\mathbf{0.8833 \pm 0.0026}$ $*$	$\mathbf{0.8921 \pm 0.0021}$	$\mathbf{0.8946 \pm 0.0022}$ $*$	$\mathbf{0.8844 \pm 0.0026}$ $*$
Breast	Single imp	0.7170 ± 0.0055	0.7200 ± 0.0048	0.7164 ± 0.0048	0.7085 ± 0.0057
	WII	$\mathbf{0.7184 \pm 0.0056}$	$\mathbf{0.7243 \pm 0.0048}$ $*$	$\mathbf{0.7212 \pm 0.0050}$ $*$	$\mathbf{0.7152 \pm 0.0057}$ $*$
Diabetes	Single imp	0.7448 ± 0.0025	0.7053 ± 0.0036	0.7154 ± 0.0043	0.7438 ± 0.0026
	WII	$\mathbf{0.7455 \pm 0.0025}$	$\mathbf{0.7234 \pm 0.0036}$ $*$	$\mathbf{0.7389 \pm 0.0031}$ $*$	$\mathbf{0.7439 \pm 0.0024}$
German	Single imp	0.7331 ± 0.0029	0.7058 ± 0.0029	0.7056 ± 0.0028	$\mathbf{0.7364 \pm 0.0029}$
	WII	$\mathbf{0.7368 \pm 0.0025}$ $*$	$\mathbf{0.7118 \pm 0.0030}$ $*$	$\mathbf{0.7120 \pm 0.0028}$ $*$	0.7357 ± 0.0027
Waveform	Single imp	0.8700 ± 0.0019	0.8241 ± 0.0031	0.7827 ± 0.0049	0.8679 ± 0.0020
	WII	0.8700 ± 0.0019	$\mathbf{0.8612 \pm 0.0019}$ $*$	$\mathbf{0.8583 \pm 0.0020}$ $*$	$\mathbf{0.8686 \pm 0.0020}$ $*$

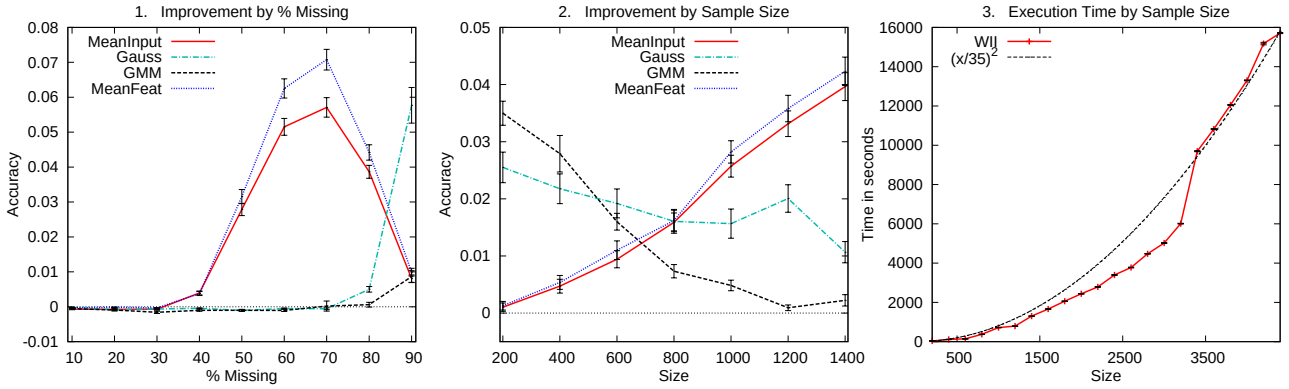

Figure 1. Detailed results on USPS classification task.

missing attributes cannot be used as test examples, the test sets are only taken from the fully observed part of the data set.

Table 6.1 shows accuracies and standard errors for the *weighted infinite imputations* (WII) method with squared distance regularization compared to all single imputations $\hat{\boldsymbol{\omega}}$ on each data set. Regularization parameter γ is automatically chosen for each run based on the performance on a separate tuning set. Baselines are obtained by first imputing $\hat{\boldsymbol{\omega}}$ and learning the classifier in a second step. The *weighted infinite imputations* method outperforms the single imputation in virtually all settings. We test for significant improvements with a paired t-test on the 5% significance level. Significant improvements are marked with a "$*$" in the table.

We explore the dependence of classification performance on training sample size and the percentage of missing attribute values in more detail. The first graph in Figure 1 shows improvements in classification accuracy of our method over the single imputations depending on the percentage of missing values. Graph 2 shows classification accuracy improvements depending on the size of the labeled training set. Both experiments are performed on USPS data set and we again adjust γ separately for each run based on the performance on the tuning set. We note that similar results are obtained for the other classification problems. The *weighted infinite imputation* method can improve classification accuracy even when only 30% of the attribute values are missing. It shows, though, that it works best if at least 60% are missing, depending on $\hat{\boldsymbol{\omega}}$. On the other hand, we see that it works for all training set sizes, again depending on $\hat{\boldsymbol{\omega}}$. Similar results are obtained for the other data sets.

Table 2. Mean squared error results and standard errors for regression data sets. Smaller mean squared errors are written in bold face, "∗" denotes significant improvement.

		MeanInput	Gauss	GMM	MeanFeat
Housing	Single imp	193.0908 ± 19.9408	288.6192 ± 41.5954	160.4940 ± 16.2004	1134.5635 ± 101.9452
	WII	$\mathbf{66.5144 \pm 0.8958}$ ∗	$\mathbf{62.3073 \pm 0.8479}$ ∗	$\mathbf{66.7959 \pm 0.9173}$ ∗	$\mathbf{64.7926 \pm 0.9619}$ ∗
Ailerons	Single imp	81.7671 ± 4.5862	172.5037 ± 8.6705	79.8924 ± 4.0297	193.5790 ± 10.4899
	WII	$\mathbf{11.8034 \pm 0.1494}$ ∗	$\mathbf{8.7505 \pm 0.0932}$ ∗	$\mathbf{11.7595 \pm 0.1530}$ ∗	$\mathbf{11.8220 \pm 0.1387}$ ∗
Cpu_act	Single imp	10454.176 ± 962.598	15000.380 ± 973.100	10123.172 ± 933.143	15710.812 ± 1099.603
	WII	$\mathbf{306.257 \pm 12.500}$ ∗	$\mathbf{204.180 \pm 5.058}$ ∗	$\mathbf{305.651 \pm 13.627}$ ∗	$\mathbf{247.988 \pm 8.010}$ ∗

To evaluate the convergence of our method, we measure classification accuracy after each iteration of the learning algorithm. It shows that classification accuracy does not change significantly after about 5 iterations for a typical γ, in this case $\gamma = 10^5$ for the USPS data set. On average the algorithm terminates after about 30-40 iterations. The computational demands of the *weighted infinite imputation* method are approximately quadratic in the training set size for the classification task, as can be seen in Graph 3 of Figure 1. This result depends on the specific risk functional R and its optimization implementation. Nevertheless, it shows that risk functionals which are solvable in quadratic time do not change their computational complexity class when learned with incomplete data.

6.2. Regression

We evaluate the *weighted infinite imputations* method on regression problems using the squared error as loss function. Consequently, risk functional R^{KRR} (Sect. 5.1) is used as R and again the squared distance regularizer Q^{sq} for Q in Optimization Problem 3. From UCI we take the Housing data (506, 14), and from the Weka homepage cpu_act (1500, 21) and ailerons (2000, 40). Ridge parameter η and RBF-kernel parameter σ were again chosen such that they lead to best results on the completely observed data. Regularization parameter γ was chosen based on the performance on a tuning set consisting of 150 examples. Results are shown in Table 2. We can see that our method outperforms the results obtained with the single imputations significantly for all settings.

7. Conclusion

We devised an optimization problem for learning decision functions from incomplete data, where the distribution p of the missing attribute values is a free parameter. The investigated method makes only minor assumptions on the distribution by the means of a regularizer on p that can be chosen freely. By simultaneously optimizing the function and the distribution of imputations, their dependency is taken into account properly. We presented a proof that the optimal solution for the joint learning problem concentrates the density mass of the distribution on finitely many imputations. This justifies the presented iterative algorithm that finds a solution. We showed that instantiations of the general learning method consistently outperform single imputations.

Acknowledgments

We gratefully acknowledge support from STRATO Rechenzentrum AG.

References

Argyriou, A., Micchelli, C., & Pontil, M. (2005). Learning convex combinations of continuously parameterized basic kernels. *Proceedings of the 18th Conference on Learning Theory.*

Chechik, G., Heitz, G., Elidan, G., Abbeel, P., & Koller, D. (2007). Max-margin classification of incomplete data. *Advances in Neural Information Processing Systems 19.*

Liao, X., Li, H., & Carin, L. (2007). Quadratically gated mixture of experts for incomplete data classification. *Proceedings of the 24th International Conference on Machine learning.*

Micchelli, C., & Pontil, M. (2007). Feature space perspectives for learning the kernel. *Machine Learning, 66.*

Shivaswamy, P. K., Bhattacharyya, C., & Smola, A. J. (2006). Second order cone programming approaches for handling missing and uncertain data. *Journal of Machine Learning Research, 7.*

Smola, A., Vishwanathan, S., & Hofmann, T. (2005). Kernel methods for missing variables. *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics.*

Williams, D., & Carin, L. (2005). Analytical kernel matrix completion with incomplete multi-view data. *Proceedings of the ICML 2005 Workshop on Learning With Multiple Views.*

Williams, D., Liao, X., Xue, Y., & Carin, L. (2005). Incomplete-data classification using logistic regression. *Proceedings of the 22nd International Conference on Machine learning.*