Understanding the use of unlabelled data in predictive modelling

Feng Liang[*], Sayan Mukherjee[*†], and Mike West[*]

feng,sayan,mike@isds.duke.edu

May 2005

Abstract

The incorporation of *unlabelled data* in statistical machine learning methods for prediction, including regression and classification, has demonstrated the potential for improved accuracy in prediction in a number of recent examples. The statistical basis for this *semi-supervised* analysis does not, however, appear to have been well delineated in the literature to date. Nor, perhaps, are statisticians as fully engaged in the vigourous research in this area of machine learning as might be desired. Much of the theoretical work in the literature has focused, for example, on geometric and structural properties of the unlabeled data in the context of particular algorithms, rather than probabilistic and statistical questions. This paper overviews the fundamental statistical foundations for predictive modelling and the general questions associated with unlabelled data, highlighting the relevance of venerable concepts of sampling design and prior specification. This theory, illustrated with a series of simple but central examples, shows precisely when, why and how unlabelled data matter.

Keywords: Bayesian analysis, Bayesian kernel regression, Predictive distribution, Semi-supervised learning, Unlabelled data.

[*]Institute of Statistics and Decision Sciences, Duke University, Durham NC 27708

[†]Institute for Genome Sciences & Policy, Duke University, Durham NC 27708

1 Introduction

Recent interest in the use of so-called unlabelled data in problems of prediction in the machine learning community has generated a growing awareness of the potential for incorporation of ancillary design data (sometimes referred to as semi-supervised learning) in classification and regression problems (Joachims, 1999; Blum and Mitchell, 1998; Szummer and Jaakkola, 2001; Zhu et al., 2003; Belkin et al., 2004; Bennett and Demiriz, 1999). Mainstream statistical thinking is relatively under-represented in this active and exciting literature; we believe, however, that statisticians have much to contribute to the emerging discussion, especially in articulation of the "when, why and how" in regard to use of unlabelled data. Machine learning examples are typically presented case-by-case, with the semi-supervised analysis usually based on modifications of (fully supervised) algorithms for classification or regression prediction, with the introduction of additional components of objective functions that tie-in unlabelled cases. Theoretical arguments for the additional components are made using a combination of structural and intuitive arguments, including, most recently, asymptotic arguments on the convergence of operators on manifolds (Lafon, 2004; Belkin, 2003). There has been some work addressing the theoretical aspect of unlabelled data (Seeger, 2001; Castelli and Cover, 1995; Cozman and Cohen, 2001; Ando and Zhang, 2004) in specific contexts. However, in general, the foundation and rationale for understanding the relevance, and likely effectiveness, of unlabelled data is still not well understood.

Beginning with an articulation of the basic model framework and assumptions of sampling and design, we discuss the underlying conceptual and theoretical basis for using unlabelled data. This is developed in the Bayesian framework for prediction, in which implications for the incorporation, or otherwise, of unlabelled data in prediction problems becomes transparent. This theoretical basis is developed in section 2, followed by a series of central, simple yet illuminating examples in section 3, and summary comments in section 4.

2 General Framework

2.1 Context, Goals and Models

Interest lies in aspects of the joint distribution of two random quantities, (y, x), and the core prediction problem concerns statements about future values of y based on observing the corresponding x. Both x and y may be multivariate, in general. In standard regression problems, y is a continuous or discrete univariate response; in problems of classification, y is discrete - often binary. Using $p(\cdot)$ as generic notation for probability density functions, all inference problems require understanding aspects of the joint density $p(y, x)$.

The fundamental problem of prediction - whether it be couched in terms of regression estimation or classification - is framed as follows: at a future specified value of x, make statements about the corresponding value of y. Using $*$ to denote future values of interest, this implies a directional focus: we want to understand $p(y_*|x_*)$ based on all available data and information.

Statistical models structure the problem in terms of parameters (which may be infinite dimensional in non-parametric models) that represent all uncertain aspects of the joint probability distribution for (y, x). By way of notation, the joint density is

$$p(y, x|\phi, \theta) = p(y|x, \phi)p(x|\theta) \tag{1}$$

where the functional forms of the two densities on the right hand side are completely specified by the characterizing parameters (ϕ, θ). Though ϕ and θ are two distinct symbols in notation, they can be dependent in various ways as we will see later in section 3. From this joint density, we can also deduce the implied marginal density for y, $p(y|\phi, \theta)$, and the implied conditional density $p(x|y, \phi, \theta)$.

Sometimes, especially in classification problems, the joint density is parametrised as

$$p(y, x|\psi, \mu) = p(x|y, \mu)p(y|\psi). \tag{2}$$

The conditional density of y given x is essential for prediction, of course, and hence we center our development on the representation (1). For a joint density parameterized as (2), we can deduce the implied marginal density for x, $p(x|\theta)$ with $\theta = \theta(\psi, \mu)$ and the implied conditional density $p(y|x, \phi)$ with $\phi = \phi(\psi, \mu)$. This is one of many examples in which the two characterising parameters θ and ϕ in (1) are functionally related in what might be rather complex ways.

2.2 Sampling Designs

We refer to the data generation process as sampling design. Data from different sampling designs provide different information about (ϕ, θ). Typical sampling contexts fall into the following categories:

1. Data from the *margins:*

   - $Y^m = \{y_1^m, \ldots, y_{k_m}^m\}$ where the $y_i^m \sim p(y|\phi, \theta)$ are independent, and/or
   - $X^m = \{x_1^m, \ldots, x_{n_m}^m\}$ where the $x_i^m \sim p(x|\theta)$ are independent,

   but with no connection whatsoever between the two. Having the opportunity to observe data Y^m provides information on aspects of the parameters (ϕ, θ), and similarly X^m informs on aspects of θ. X^m is the traditional *unlabelled data,* though we see that the same term could also be applied to Y^m.

2. Data from a *prospective design*: $(Y^p, X^p) = \{(y_i^p, x_i^p);\ i = 1, \ldots, n_p\}$ are drawn as a random sample from the full joint distribution $p(y, x|\phi, \theta)$. Here data are paired and provide information on both θ and ϕ. This is the usual regression or classification design. However, in some cases the $X^p = \{x_1^p, \ldots, x_{n_p}^p\}$ values are pre-fixed, i.e., specified in advance by design. In such a case X^p contains no information about θ, and we learn about the parameter ϕ through the likelihood comprised of components $p(y_i^p|x_i^p, \phi)$. An interesting example of this design in machine learning is the transductive framework, outlined by Vapnik (1998), where the objective is to make predictions on only pre-specified values of x.
3. Data from a typical *retrospective design* (or *case-control design*): we observe the outcomes $X^r = \{x_1^r, \ldots, x_{n_r}^r\}$ at a chosen set of y values $Y^r = \{y_1^r, \ldots, y_{n_r}^r\}$. Here the data are paired, but Y^r provide no information about (ϕ, θ) since the y values are chosen by design. The data in X^r comprise a set of n_r independent random draws from $p(x|y, \phi, \theta)$ and therefore provide information about (ϕ, θ).

The difference between "prospective" and "retrospective" is whether the observed y values are random or not. Since most examples we will discuss come from a prospective design, for notational simplicity we will drop the superscript and use (Y, X) to denote (Y^p, X^p).

In standard machine learning problems the term "sampling" generally does not relate to the data generation mechanism, but to different parameterizations (or factorizations) of the joint distribution assuming the data was generated by a prospective design with x, y random: the parameterization (1) is referred to as "diagnostic sampling" and (2) is referred to as "generative sampling" (Seeger, 2001; Cozman and Cohen, 2001).

2.3 Prediction

Suppose we observe data D that provide information about (ϕ, θ) summarized in terms of the implied posterior distribution with density $p(\phi, \theta|D)$. We aim to predict (estimate, classify) a new case y_* at a fixed (specified, chosen) value x_*. The prediction problem is solved from the Bayesian perspective by evaluating the predictive distribution

$$p(y_*|x_*, D) = \int_{\phi \in \Phi, \theta \in \Theta} p(y_*|x_*, \phi) p(\phi, \theta|x_*, D) d\phi d\theta$$

at the assumed value of the future x_*. This is the key to understanding if - and, if so, how - any information in D, including unlabelled observations of any kind, impacts on the prediction problem.

In some cases x_* will arise as a sample from $p(x|\theta)$ and so provide information about θ. In this case $p(\phi, \theta|x_*, D)$ depends on x_*. In other cases x_* is chosen at values we aim to explore

for potential future outcomes, so that

$$p(\phi, \theta | x_*, D) = p(\phi, \theta | D). \tag{3}$$

In any example it is important to be aware of the distinction but, for our development here, it is a side issue and we assume the latter case (3) as it simplifies the notation.

The usual notion is that X^m is the unlabelled data in question - information relevant to understanding the distribution of the predictor variables alone. Hence interest focuses here on how X^m comes into the evaluation of the above predictive density. All forms of information enter in through D, so for X^m (and any other information) to be relevant in prediction it is necessary that it play a role in defining the posterior $p(\phi, \theta | D)$.

Regarding the definition of each of the possible data sources arising, the most general framework has observations on each of $Y^m, X^m, (Y, X), (Y^r, X^r)$. In this most general case we then have, via Bayes' theorem and under a specified prior $p(\phi, \theta)$,

$$p(\phi, \theta | D) \propto p(\phi, \theta) p(D | \phi, \theta)$$

with

$$p(D | \phi, \theta) = p(Y, X | \phi, \theta) p(X^m | \theta) p(Y^m | \phi, \theta) p(X^r | Y^r, \phi, \theta).$$

In general, this will depend in complicated ways on all aspects of D, including various aspects of the unlabelled data X^m. Investigating this dependence is the key to understanding the relevance and specific potential uses of unlabelled data . Some specific and typical cases focus the discussion.

2.4 Common Framework of Regression and Classification

For convenience and clarity, we start our discussion in the simple regression/classification context where data arise from a joint random sample $D = (X, Y)$. Then

$$p(\phi, \theta | D) \propto p(\phi, \theta) p(Y | X, \phi) p(X | \theta).$$

For example, we may have a linear or nonlinear regression model for $(y|x)$ in which ϕ represents the uncertain regression parameters or regression function.

Now imagine that we have the opportunity to additionally observe or measure some unlabelled data X^m. The modified posterior with $D = \{Y, X, X^m\}$ is then

$$p(\phi, \theta | D) \propto p(\phi, \theta) p(Y | X, \phi) p(X | \theta) p(X^m | \theta).$$

We deduce the following:

- If ϕ and θ are independent under the prior, $p(\phi, \theta) = p(\phi)p(\theta)$, then

$$p(\phi, \theta|D) = p(\phi|Y, X)p(\theta|X^m, X).$$

  That is, prior independence leads to posterior independence, and, as a result
- the unlabelled data X^m is irrelevant to learning about ϕ, and hence irrelevant in predicting new y_*, if ϕ and θ are *a priorí* independent. This is because

$$p(y_*|x_*, D) = \int p(y_*|x_*, \phi)p(\phi, \theta|D)d\phi d\theta = \int p(y_*|x_*, \phi)p(\phi|Y, X)d\phi$$

  by the posterior independence.

In other cases, the posterior for (θ, ϕ) may involve dependencies. Therefore, additional information generated from marginal data will have an impact on the prediction problem via the integration over the posterior that defines $p(y_*|x_*, D)$. In the general framework, data from Y^m, X^m, and (Y^r, X^r) will all have an impact on the prediction problem.

In this simple context, it is transparent that formal model-prior connections between the "regression component" parameters ϕ and the "$x-$marginal" component parameters θ are necessary if unlabelled data, as it is traditionally defined, is to play a role. How such dependencies arise, and what forms they take, depend on context, and some examples now illuminate this.

3 Specific Contexts and Examples

3.1 Normal linear regression models

In the usual normal linear regression model, $\phi = (\beta, \tau)$ is the set of regression parameters from the model

$$y|x, \phi \sim N(\beta' x, \tau^2)$$

where x and β are k-dimensional vectors. One way such models arise is from assumed joint multivariate normal distributions for (y, x'), namely the $(k+1)-$dimensional normal $N(\mu, \Sigma)$ where

$$\mu = \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \ \Sigma = \begin{pmatrix} \sigma_y^2 & \rho' \\ \rho & \Sigma_x \end{pmatrix},$$

for some scalar parameters μ_y, σ_y, $k-$dimensional vector of covariance parameters ρ, and $k \times k$ variance matrix Σ_x. Under such a model we have $p(x|\theta) = N(\mu_x, \Sigma_x)$ and characterizing parameter sets $\phi = (\beta, \tau)$ and $\theta = (\mu_x, \Sigma_x)$, each being obtained via parameter transformations from (μ, Σ).

Some contexts include:

- A direct specification of the prior $p(\theta, \phi) = p(\theta)p(\phi)$ that assumes independence, and so implies that unlabelled X^m data will be irrelevant to prediction of future y_*.
- A direct specification of the prior $p(\theta, \phi)$ with dependence, such as $\beta|\phi \sim N(0, \tau^2\Sigma_x^{-1})$, which induces a relevance of the unlabelled data since X^m provides information about ϕ *indirectly* through its relevance for θ.
- An indirect specification in which the initial prior is defined for (μ, Σ), with the prior $p(\phi, \theta)$ being implied by transformation. A common approach is to use the conjugate normal-inverse Wishart prior distribution. Any prior in this class has the property that the implied prior on (ϕ, θ) is in fact one in which ϕ and θ are independent (Geiger and Heckerman, 2002; Dobra et al., 2004).

This last example illustrates a case in which modelling prior information on parameters of the joint distribution of y and x using a standard conjugate implies that the unlabelled X^m data will be irrelevant for predicting y_*. This result arises more generally in exponential family models. Other priors may, and usually will, lead to prior and then to posterior dependence, and in such cases, then, unlabelled data is relevant.

3.2 Binary outcomes example

A simple but illuminating example is the case of binary y and binary x. For thematic context, suppose $x = 1/0$ represents the presence/absence of mutation in the BRCA1 breast cancer gene in a woman, and that $y = 1/0$ represents occurrence of breast cancer before age 70.

In terms of the directional specification of the joint density as $p(y|x, \phi)p(x|\theta)$, the parameters are now just three probabilities, $\phi = (\phi_0, \phi_1)$ and θ where

- $\phi_x = p(y = 1|x, \phi)$ for $x = \{0, 1\}$, and
- $\theta = p(x = 1|\theta)$.

In the breast cancer genetics example, θ is the incidence rate of the BRCA1 mutation, ϕ_0 is the base rate for breast cancer in the general (wild type) population of women and θ_1 the (higher) cancer rate among carriers of the mutation.

Here we have prediction defined by

$$p_* = p(y_* = 1|x_*, D) = \int_{\Phi,\Theta} \phi_{x_*} p(\phi, \theta|x_*, D) d\phi d\theta.$$

Then:

- As in regression example above, if we directly specify independent priors on θ and ϕ then p_* does not depend on the unlabelled data.

- In the joint space, we have cell probabilities $p(x, y)$ on $x = 0, 1, y = 0, 1$ defined by

$$\pi = \{\pi_{0,0}, \pi_{0,1}, \pi_{1,0}, \pi_{1,1}\}.$$

  Common approaches utilize Dirichlet priors on π. If we choose a Dirichlet prior $p(\pi)$ and find the implied prior $p(\phi, \theta)$ by transformation, the result is prior independence of ϕ and θ, and again the unlabelled data is irrelevant to prediction.
- Suppose we have a prior on π that is a discrete mixture of two (or more) Dirichlets. For example, suppose that our breast cancer samples come from an inhomogeneous population having two genetically and environmentally different subpopulations in connection with inherited breast cancer related characteristics and lifetime cancer risks. In this case a reasonable prior would have the form

$$p(\pi) = ap_0(\pi) + (1 - a)p_1(\pi)$$

  where p_0 and p_1 are two different Dirichlet priors, though the sampling design cannot distinguish between the subpopulations.

  It then follows by transformation that

$$p(\phi|\theta) = w(\theta)p_0(\phi) + (1 - w(\theta))p_1(\phi)$$

  where p_0 and p_1 are the implied margins on ϕ from each of the two Dirichlets, and the mixing probability $w(\theta)$ is computed, at any conditioning value of θ, using

$$\frac{w(\theta)}{1 - w(\theta)} = \frac{a}{(1 - a)} \frac{p_0(\theta)}{(1 - p_1(\theta))}.$$

  Thus under a mixture prior of this form, θ and ϕ are dependent. Hence the unlabelled data X^m will feed through to provide information about y_* indirectly via θ and the ϕ (unless, of course, $p_0(\phi)$ and $p_1(\phi)$ happen to be equal).

3.3 Mixture model examples

In classification problems, it is often assumed that x is from a mixture distribution and each component of the mixture corresponds to a class which is indicated by y (West, 1992). For example, in binary classification, the joint distribution can be described as follows: $x \sim f_0(x)$ when $y = 0$, $x \sim f_1(x)$ when $y = 1$, and $p(y = 1) = \pi$ where $0 \leq \pi \leq 1$. In a Gaussian mixture model, for example, f_0 and f_1 are Gaussian densities parameterized by different mean and covariance structure (Lavine and West, 1992). Here we parameterize the joint density by the marginal of y and the conditional of x given y. If we then transform it back to the (ϕ, θ) parameterization as in (1), we see that θ and ϕ are dependent as we discussed at the end of

section 2.1. So it is clear that the unlabelled data X^m is informative in this context. Indeed, various semi-supervised approaches have been proposed to take advantage of the information in X^m and the effectiveness of X^m has been either implicit or explicitly well-studied (Mueller et al., 1996; Nigam et al., 1998; O'Neill, 1978; Ganesalingam and McLachlan, 1979, 1978; Castelli and Cover, 1995).

Such a model also arises in retrospective studies. Since y values are pre-specified there, the likelihood of (x_i^r, y_i^r) is equal to either f_0 or f_1 and does not depend on π; information about π is, for example, generated from observations on unlabelled Y^m data. In this case, unlabelled data X^m or Y^m is not only informative and relevant, but is also necessary to the prediction problem.

3.4 Factor regression example

Factor regression is a useful tool for regression problems with high dimensional predictors. Regression of a response variable on principal components (singular factors) is a special limiting case of *"empirical factor model"*.

West (2003) formalised the development of large-scale, latent factor models coupled with regression on latent factors, and so delineated a comprehensive framework for predictive modelling in the *"large p, small n"* paradigm. This elucidated the theory underlying Bayesian modelling using principal component projections of high-dimensional covariates/predictors as a limiting case of a broader class of regression models where the predictors are *latent* variables. This framework and theory also clarified and justified the use of so-called $g-$priors (Zellner, 1986) for Bayesian shrinkage regression, and defined novel classes of multiple shrinkage methods that are significantly beneficial in prediction problems through the ability to induce differential *shrinkage* in different factor-predictor dimensions.

To be specific in the context of a normal linear model example (the principles are of course more general), suppose we have a model in which a univariate (for the sake of example) response y is to be predicted based on a (high-dimensional) $p \times 1$ predictor variable x, and we have

$$y_i = \alpha' \lambda_i + \epsilon_i$$

and

$$x_i = B\lambda_i + \nu_i$$

where $\epsilon_i \sim N(0, \sigma^2)$, $\lambda_i \sim N(0, I)$ is a $k \times 1$ unit multivariate normal latent factor for each i, B is an uncertain $p \times k$ matrix of factor loadings of x on λ, $\nu_i \sim N(0, \Psi)$ is a vector of idiosyncratic noise term, and Ψ is an uncertain diagonal variance matrix. Also, the ν_i and ϵ_i are conditionally (on all model parameters) mutually independent and independent over i.

This framework is a key example of context when unlabelled data matter. Fundamentally, the outcomes y to be predicted are modelled as responses in regressions on *latent* variables λ, and the *observed* concomitant x variables are related to λ, while y and x are conditionally independent *given* λ. Thus the predictive relevance of x is indirect, through λ.

By marginalization over λ of the implied joint multivariate normal distribution of y, x and λ, it becomes clear that we can identify $p(y|x, \phi)$ as a normal linear regression of y on x with regression parameters and residual variance $\phi = \phi(\alpha, \sigma, B, \Psi)$. Also, the implied margin for x is normal with zero mean and variance matrix $\theta = BB' + \Psi$. In this framework, if $\{B, \Psi\}$ are known, then θ is known. So the values of any observed, unlabelled data X^m has no influence whatsoever in the problem of predicting a future y_* given other data from either prospective or retrospective designs. However, the usual circumstance is that in which $\{B, \Psi\}$ are uncertain and to be estimated. Then,

- Additional unlabelled data X^m provide information relevant to improved estimation of these parameters, and hence of relevance to predicting future y_* values via the transfer of information through inferences on the future λ_* related to x_*.
- The dependence of ϕ on aspects of θ, indirectly through their functional associations with the factor model parameters, implies that any relevant prior $p(B, \Psi, \alpha, \sigma)$ will induce prior dependencies between ϕ and θ.

3.5 Kernel regression example

An interesting class of examples, which are central to the methodological interfaces of statistics and machine learning, arise in models based on kernel regression, including Bayesian models of kernel regression (Liao et al., 2005).

The context is non-parametric, non-linear regression with $y \in \mathbb{R}$, $x \in \mathbb{R}^k$, and a model of the form

$$y = f(x) + \epsilon,$$

where ϵ is zero-mean noise term and $f(\cdot)$ is an uncertain regression function. As an example, the class of Bayesian radial basis (RB) models (Liao et al., 2005) deals with questions of proper probability models - and the resulting proper inference and predictive results that then arise - for uncertain knots in a kernel model. This framework, and other approaches, begins with the interest in a representation of the form

$$f(x) = \int w(u)k(x, u)dG(u) \tag{4}$$

for some weight function $w(u)$ over $k-$dimensional u, and some specified kernel function $k(\cdot, \cdot)$. The element $G(\cdot)$ is a probability distribution function in $k-$dimension. The key to the

model is to note that, if G is discrete and put masses g_i at support points (or "knots") u_i, then the expression for $f(\cdot)$ is simply

$$f(x) = \sum_i g_i w(u_i) k(x, u_i),$$

i.e., a radial basis function representation. The analysis of Liao et al. (2005) describes approximations to a model in which uncertainty about G is expressed using a Dirichlet process prior (Ferguson, 1973; Escobar and West, 1995). One implication of such a model for G is that, since Dirichlet processes are discrete with probability one, the formal mathematical model for $f(x)$ is the sum above with a countably infinite number of knots u_i. From the methodological viewpoint, both labelled and unlabelled x values provide information about G directly. In fact, with a sample of n labelled and/or labelled x values $x_1, \ldots, x_n$ (whether from X, X^m, or some combination of the two), this Dirichlet process model implies that f may be approximated by

$$\hat{f}_n(x) = \sum_{i=1}^{n} w_{n,i} k(x, x_i) \tag{5}$$

where $w_{n,i} \propto w(x_i)$. The key methodological relevance of this approach is that this is true for all n, so providing consistency as sample size increases and additional design points are observed. This leads to the practical model in which each y_* is linearly regressed on the set of kernel predictors $\{k(x_*, x_i)\}$ based on whatever set of design points are observed.

This is a perfect example of when, why and how unlabelled X^m data matter, and of course the conclusions hold for other versions of kernel and RB analysis. In particular, we note that:

- $\theta = G(\cdot)$ so that $p(x|\theta)dx = dG(x)$ - the parameter is the full distribution function itself.
- $p(y|x, \phi)$ depends intimately on $\theta = G$ as defining the nonlinear kernel regression; in fact, $\theta \subseteq \phi$ in this case. Thus prior and posterior dependence of θ and ϕ is central to the model.
- As a result, unlabelled X^m provides direct, immediately and hugely relevant information in prediction of y_*.

Key connections with machine learning approaches are made by noting that the central model of equation (4) corresponds to the solution of the manifold regularization formulation of Belkin et al. (2004). This approach, motivated by geometric arguments, is an optimization that seeks

$$f_* = \arg \min_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^{n} V(f(x_i), y_i) + \gamma_A ||f||_K^2 + \gamma_I ||f||_I^2,$$

where $\{(y_i, x_i)\}_{i=1}^n$ are the labelled data, $\mathcal{H}_K$ is a Reproducing Kernel Hilbert Space (RKHS), $V(f(x), y)$ is a loss function, $||f||_K^2$ is the RKHS norm, γ_A, γ_I are regularization parameters, and $||f||_I^2$ is an norm that reflects the smoothness of the function on the marginal $p(x)$. If

the marginal is concentrated on a manifold, $x \subset \mathcal{M} \in \mathbb{R}^k$, then a natural choice for $||f||_I^2$ is the Laplacian on the manifold. In general, the marginal $p(x)$ is not given but we may have unlabelled data X^m from the marginal, in which case the Laplacian on the manifold may be approximated by a Laplacian on the graph defined by the observed data (labelled and unlabelled)

$$\hat{f}_n(x) = \arg\min_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^{n} V(f(x_i), y_i) + \gamma_A ||f||_K^2 + \frac{\gamma_I}{(n+n_m)^2} \mathbf{f}^T \mathbf{L} \mathbf{f},$$

where $\mathbf{L}$ is the graph Laplacian on all the data (given a weight matrix on the graph) and $\mathbf{f} = \{f(x_1), ..., f(x_n), f(x_1^m), ..., f(x_{n_m}^m)\}$. The solution to the above optimization problem has the form

$$\hat{f}(x) = \sum_{i=1}^{n} w_i k(x, x_i) + \sum_{i=1}^{n_m} w_{n+i} k(x, x_i^m),$$

which takes the same form as approximation (5) formulated from the Dirichlet process prior.

4 Summary comments

Beginning with a simple, clear articulation of the basic sampling and design specifications underlying statistical formulations of prediction problems, we have delineated the theoretical issues underlying the use and relevance, or irrelevance, of unlabelled data in classification and prediction problems. This, coupled with a series of central yet simply described and understood examples, provides an overview and synthesis of the ideas underlying the emerging methodology of semi-supervised learning in the machine learning and statistics literatures.

Graphical model representations of the joint sampling model context aid in this interpretation. The relevance, or otherwise, of the unlabelled X^m data can be deduced essentially by inspection of the implied (undirected) graphical representation of any full model structure. For example, the full distribution assuming joint sampling, and in cases for which $p(\phi, \theta) = p(\phi)p(\theta)$, is illustrated in graphical terms in Figure 1. The joint density exhibited here is

$$p(y_*, x_*, Y, X, X^m, \phi, \theta) = p(y_*|x_*, \phi)p(Y|X, \phi)p(X|\theta)p(x_*|\theta)p(X^m|\theta)p(\phi)p(\theta).$$

Figure 1(a) is a directed acyclic graph of the joint distribution structured in terms of composition of sampling distributions. Figure 1(b) displays the corresponding undirected graph in which the lack of an edge between X^m and y_* indicates conditional independence given all other quantities, hence the irrelevance to prediction of the unlabelled data in this case. In contrast, were ϕ, θ to be *a priorí* dependent, then the five nodes of the undirected graph would be fully connected, exhibiting the relevance of the unlabelled data to prediction of y_*.

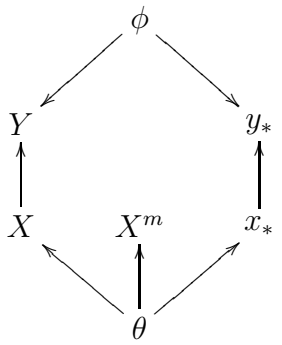

Figure 1(a)

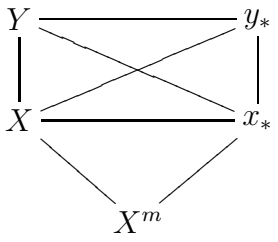

Figure 1(b)

Figure 1: Directed graph (a) and undirected graph (b) in cases of independence of ϕ, θ.

In addition to clarifying and exemplifying the structure of models and the prediction problem with unlabelled data, one aim of this work has been to review the area to provide a link across the mainstream statistical to machine learning communities. We hope that this will entice more statistical researchers into a very active, productive and exiting research milieu, while also founding the discussion in venerable, simple and unambiguous terms arising from the direct and classical probabilistic formulation. This view directly, we believe, addresses and answers the questions of "when, why and how" unlabelled data help in predictive modelling.

Acknowledgments

The authors acknowledge support of the National Science Foundation (grants DMS-0342172 and DMS 0406115). Any opinions, findings and conclusions or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of the NSF.

References

Ando, R. K. and T. Zhang (2004). A framework for learning predictive structures from multiple tasks and unlabeled data. Technical Report RC23462, IBM T.J. Watson Research Center.

Belkin, M. (2003). *Problems of Learning on Manifolds.* Ph. D. thesis, U. Chicago, Hyde Park, IL.

Belkin, M., P. Niyugi, and V. Sindhwani (2004). Manifold regularization: A geometric framework for learning from examples. Technical Report TR-2004-06, University of Chicago.

Bennett, K. P. and A. Demiriz (1999). Semi-supervised support vector machines. In *Proceedings of the 1998 conference on Advances in neural informati on processing systems II*, pp. 368–374. MIT Press.

Blum, A. and T. Mitchell (1998). Combining labeled and unlabeled data with co-training. In *COLT: Proceedings of the Workshop on Computational Learning Theory*. Morgan Kaufmann Publishers.

Castelli, V. and T. Cover (1995). On the exponential value of labeled samples. *Pattern Recognition Letters 16*(1), 105–111.

Cozman, F. G. and I. Cohen (2002). Unlabeled data can degrade classification performance of generative classifiers. In *Fifteenth International Florida Artificial Intelligence Society Conference*, pp. 327–331.

Dobra, A., B. Jones, C. Hans, J. Nevins, and M. West (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis 90*, 196–212.

Escobar, M. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association 90*, 577–588.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics 1*, 209–230.

Ganesalingam, S. and G. J. McLachlan (1978). The efficiency of a linear discriminant function based on unclassified initial samples. *Biometrika*, 658–662.

Ganesalingam, S. and G. J. McLachlan (1979). Small sample results for a linear discriminant function estimated from a mixture of normal populations. *Journal of Statistical Computation and Simulation*, 151–158.

Geiger, D. and D. Heckerman (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Annals of Statistics 5*, 1412–1440.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. In *Proceedings of ICML-99, 16th International Conference on Machine Learning*, pp. 200–209. Morgan Kaufmann Publishers.

Lafon, S. (2004). *Diffusion Maps and Geodesic Harmonics*. Ph. D. thesis, Yale, New Haven, CT.

Lavine, M. and M. West (1992). A bayesian method for classification and discrimination. *Canadian Journal of Statistics 20*, 451–461.

Liao, M., F. Liang, S. Mukherjee, and M. West (2005). Bayesian kernel regression and radial basis function models. ISDS Discussion Paper 2005, forthcoming.

Mueller, P., A. Erkanli, and M. West (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika 83*, 67–79.

Nigam, K., A. K. McCallum, S. Thrun, and T. M. Mitchell (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning 39*(2/3), 103–134.

O'Neill, T. J. (1978). Normal discrimination with unclassified observations. *Journal of the American Statistical Association*, 821–826.

Seeger, M. (2001). Input-dependent regularization of conditional density models. Technical report, University of Edinburgh.

Szummer, M. and T. Jaakkola (2001). Partially labeled classification with markov random walks. In *Advances in Neural Information Processing Systems (NIPS)*, Volume 14.

Vapnik, V. N. (1998). *Statistical Learning Theory*. New York: Wiley.

West, M. (1992). Modelling with mixtures (with discussion). In J. B. et al. (Ed.), *Bayesian Statistics 4*. Oxford University Press.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In J. B. et al. (Ed.), *Bayesian Statistics 7*, pp. 723–732. Oxford University Press.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pp. 233–243. North-Holland/Elsevier.

Zhu, X., Z. Ghahramani, and J. Lafferty (2003). Semi-supervised learning using gaussian fields and harmonic functions. In *The Twentieth International Conference on Machine Learning (ICML-2003)*, Volume 20.