Models and Selection Criteria for Regression and Classification

David Heckerman
heckerma@microsoft.com

Christopher Meek
meek@microsoft.com

May 1997

Technical Report
MSR-TR-97-08

Microsoft Research
Advanced Technology Division
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

Abstract

When performing regression or classification, we are interested in the conditional probability distribution for an *outcome* or *class* variable Y given a set of *explanatory* or *input* variables $\mathbf{X}$. We consider Bayesian models for this task. In particular, we examine a special class of models, which we call Bayesian regression/classification (BRC) models, that can be factored into independent conditional $(y|\mathbf{x})$ and input $(\mathbf{x})$ models. These models are convenient, because the conditional model (the portion of the full model that we care about) can be analyzed by itself. We examine the practice of transforming arbitrary Bayesian models to BRC models, and argue that this practice is often inappropriate because it ignores prior knowledge that may be important for learning. In addition, we examine Bayesian methods for learning models from data. We discuss two criteria for Bayesian model selection that are appropriate for repression/classification: one described by Spiegelhalter et al. (1993), and another by Buntine (1993). We contrast these two criteria using the prequential framework of Dawid (1984), and give sufficient conditions under which the criteria agree.

Keywords: Bayesian networks, regression, classification, model averaging, model selection, prequential criteria

1 Introduction

Most work on learning Bayesian networks from data has concentrated on the determination of relationships among a set of variables. This task, which we call *joint analysis*[1], has applications in causal discovery and the prediction of a set of observations. Another important task is *regression/classification*: the determination of a conditional probability distribution for an *outcome* or *class* variable Y given a set of *explanatory* or *input* variables $\mathbf{X}$. When Y has a finite number of states we refer to the task as *classification*. Otherwise we refer to the task as *regression*.

In this paper, we examine parametric models for the regression/classification task. In Section 2, we examine a special class of models, which we call Bayesian regression/classification (BRC) models, that can be factored into independent conditional $(y|\mathbf{x})$ and input $(\mathbf{x})$ models. These models are convenient, because the conditional model (the portion of the full model that we care about) can be analyzed alone. In Section 3, we examine the practice of transforming arbitrary Bayesian models to BRC models, and argue that this practice is often inappropriate because it ignores prior knowledge that may be important for learning.

Also in this paper, we discuss Bayesian methods for learning models from data. In Section 4, we compare Bayesian model averaging and model selection. In Section 5, we discuss

[1] This task is sometimes called *density estimation*.

two criteria for Bayesian model selection that are appropriate for regression/classification: one described by Spiegelhalter et al. (1993), and another by Buntine (1993). We contrast these two criteria using the prequential framework of Dawid (1984), and give sufficient conditions under which the criteria agree.

The terminology and notation we need is as follows. We denote a variable by an upper-case letter (e.g., X, Y, X_i, Θ), and the state or value of a corresponding variable by that same letter in lower case (e.g., x, y, x_i, θ). We denote a set of variables by a bold-face upper-case letter (e.g., $\mathbf{X}, \mathbf{Y}, \mathbf{X}_i$). We use a corresponding bold-face lower-case letter (e.g., $\mathbf{x}, \mathbf{y}, \mathbf{x}_i$) to denote an assignment of state or value to each variable in a given set. We say that variable set $\mathbf{X}$ is in *configuration* $\mathbf{x}$. We use $p(\mathbf{X} = \mathbf{x}|\mathbf{Y} = \mathbf{y})$ (or $p(\mathbf{x}|\mathbf{y})$ as a shorthand) to denote the probability or probability density that $\mathbf{X} = \mathbf{x}$ given $\mathbf{Y} = \mathbf{y}$. We also use $p(\mathbf{x}|\mathbf{y})$ to denote the probability distribution (both mass functions and density functions) for $\mathbf{X}$ given $\mathbf{Y} = \mathbf{y}$. Whether $p(\mathbf{x}|\mathbf{y})$ refers to a probability, a probability density, or a probability distribution will be clear from context.

We use $\mathbf{m}$ and $\boldsymbol{\theta}_m$ to denote the structure and parameters of a model, respectively. When $(\mathbf{m}, \boldsymbol{\theta}_m)$ is a Bayesian network for variables $\mathbf{Z}$, we write the usual factorization as

$$p(z_1, \ldots, z_N|\boldsymbol{\theta}_m, \mathbf{m}) = \prod_{i=1}^{N} p(z_i|\mathbf{pa}_i, \boldsymbol{\theta}_m, \mathbf{m}) \tag{1}$$

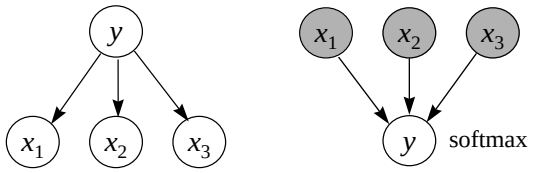
where $\mathbf{Pa}_i$ are the variables corresponding to the parents of Z_i in $\mathbf{m}$. We refer to $p(z_i|\mathbf{pa}_i, \boldsymbol{\theta}_m, \mathbf{m})$ as the *local distribution function* for Z_i.

2 Models for Regression/Classification

In this section, we examine various parametric models for the task of regression/classification. Models for this task are of two main types: conditional models and joint models. A *conditional model* is of the form $p(y|\mathbf{x}, \boldsymbol{\theta}_m, \mathbf{m})$. A *joint model* is of the form $p(y, \mathbf{x}|\boldsymbol{\theta}_m, \mathbf{m})$. We use a joint model for regression/classification by performing probabilistic inference to obtain $p(y|\mathbf{x}, \boldsymbol{\theta}_m, \mathbf{m})$.

Examples of joint models include Bayesian networks. Figure 1a shows the structure of a *naive Bayes* model in which the variables $\mathbf{X}$ are mutually independent given Y. Suppose Y has r states $y^1, \ldots, y^r$, each X_i is binary with states x_i^1 and x_i^2, and each local distribution function is a collection of multinomial distributions (one distribution for each parent configuration). For this example, it is not difficult to derive the corresponding conditional model (see, for example, Bishop, 1995, Chapter 6). Namely, we have

$$\lambda_{k\mathbf{x}} \equiv \log \frac{p(y^k|\mathbf{x}, \boldsymbol{\theta}_m, \mathbf{m})}{p(y^1|\mathbf{x}, \boldsymbol{\theta}_m, \mathbf{m})} = \log \frac{\theta(y^k)}{\theta(y^1)} + \sum_{i=1}^{n} \log \frac{\theta(x_i|y^k)}{\theta(x_i|y^1)} \tag{2}$$



(a) (b)

Figure 1: (a) A naive Bayes model for classification. (b) A linear softmax regression model that has the same conditional distribution for Y.

for $k = 2, \ldots, r$. After some algebra, Equation 2 becomes

$$\lambda_{k\mathbf{x}} = \left(\log \frac{\theta(y^k)}{\theta(y^1)} + \sum_{i=1}^{n} \log \frac{\theta(x_i^1|y^k)}{\theta(x_i^1|y^1)} \right) + \sum_{i=1}^{n} I(x_i^2) \left(\frac{\theta(x_i^2|y^k)}{\theta(x_i^2|y^1)} - \frac{\theta(x_i^1|y^k)}{\theta(x_i^1|y^1)} \right) \tag{3}$$

where $I(x_i^2)$ is the indicator variable that is equal to 1 if and only if $x_i = x_i^2$. Consequently, we have

$$p(y^k|\mathbf{x}, \boldsymbol{\theta}_m, \mathbf{m}) = \frac{e^{\lambda_{k\mathbf{x}}}}{1 + \sum_{j=2}^{r} e^{\lambda_{j\mathbf{x}}}} \equiv \text{softmax}(\lambda_{1\mathbf{x}}, \ldots, \lambda_{k\mathbf{x}}) \tag{4}$$

where each $\lambda_{k\mathbf{x}}$ is a linear function of $I(x_1^2), \ldots, I(x_n^2)$.

This conditional model $p(y|\mathbf{x}, \boldsymbol{\theta}_m, \mathbf{m})$ is a type of generalized linear model known as a *linear softmax regression*.[2] We can display the structure of this conditional model as a Bayesian network, as shown in Figure 1b. In the figure, the input nodes $\mathbf{X}$ are shaded to indicate that we observe them and hence do not care about their joint distribution.

Now let us specialize our discussion to Bayesian models for regression/classification. In the Bayesian approach, we encode our uncertainty about $\boldsymbol{\theta}_m$ and $\mathbf{m}$ using probability distributions $p(\boldsymbol{\theta}_m|\mathbf{m})$ and $p(\mathbf{m})$, respectively. Thus, the Bayesian variant of a joint model takes the form

$$p(y, \mathbf{x}, \boldsymbol{\theta}_m, \mathbf{m}) = p(\mathbf{m})\ p(\boldsymbol{\theta}_m|\mathbf{m})\ p(y, \mathbf{x}|\boldsymbol{\theta}_m, \mathbf{m}) \tag{5}$$

We refer to this model as a *Bayesian joint* (BJ) model.

We define a Bayesian analogue to a conditional model as follows. Suppose that $\boldsymbol{\theta}_m$ can be decomposed into parameters $(\boldsymbol{\theta}_{\mathbf{x}}, \boldsymbol{\theta}_{y|\mathbf{x}})$ such that

$$p(y, \mathbf{x}|\boldsymbol{\theta}_m, \mathbf{m}) = p(\mathbf{x}|\boldsymbol{\theta}_{\mathbf{x}}, \mathbf{m})\ p(y|\mathbf{x}, \boldsymbol{\theta}_{y|\mathbf{x}}, \mathbf{m}) \tag{6}$$

$$p(\boldsymbol{\theta}_m|\mathbf{m}) = p(\boldsymbol{\theta}_{\mathbf{x}}|\mathbf{m})\ p(\boldsymbol{\theta}_{y|\mathbf{x}}|\mathbf{m}) \tag{7}$$

[2] Although Y has a finite number of states, this model is commonly referred to as a regression.

In this case, given data $D = ((y_1, \mathbf{x}_1), \ldots, (y_N, \mathbf{x}_N))$, assumed to be a random sample from the true distribution of Y and $\mathbf{X}$, we have

$$p(\boldsymbol{\theta}_m | y, \mathbf{x}, \mathbf{m}) \propto \{p(\mathbf{x}|\boldsymbol{\theta}_\mathbf{x}, \mathbf{m}) p(\boldsymbol{\theta}_\mathbf{x}|\mathbf{m})\} \cdot \left\{p(y|\mathbf{x}, \boldsymbol{\theta}_{y|\mathbf{x}}, \mathbf{m}) p(\boldsymbol{\theta}_{y|\mathbf{x}}|\mathbf{m})\right\}$$

Consequently, we can analyze the marginal ($\mathbf{x}$) and conditional ($y|\mathbf{x}$) terms independently. In particular, if we care only about the conditional distribution, we can analyze it on its own. We call this model defined by Equations 6 and 7 a *Bayesian regression/classification* (BRC) model. Simple examples of BRC models include ordinary linear regression (e.g., Gelman et al., 1995, Chapter8), and generalized linear models (e.g., Bishop, 1995, Chapter 10).

Note that our Bayesian analogue to the conditional model is a special case of a BJ model. One could imagine using a Bayesian model that encodes only the conditional likelihood $p(y|\mathbf{x}, \boldsymbol{\theta}_m, \mathbf{m})$ and a joint distribution for $\boldsymbol{\theta}_m$ and $\mathbf{m}$. However, this approach is flawed, because it may miss important relationships among the domain variables or their parameters that are important for learning. In the following section, we consider an example of this point.

3 Embedded Regression/Classification Models

A special class of BRC models is suggested by the following observation. For many BJ models, the conditional likelihood $p(y|\mathbf{x}, \boldsymbol{\theta}_m, \mathbf{m})$ is a simple function of $\mathbf{x}$, whereas the expression for the input likelihood $p(\mathbf{x}|\boldsymbol{\theta}_m, \mathbf{m})$ is more complicated. For example, given a naive-Bayes model in which the variables $\mathbf{X}$ are mutually independent given Y, the conditional likelihood is a simple generalized linear model, but the input likelihood is a mixture distribution. Thus, assuming we are interested in the task of regression/classification, we can imagine extracting the conditional likelihood from a BJ model, and embedding it in a BRC model. In particular, given a BJ model $(\boldsymbol{\theta}_m, \mathbf{m})$, we can create a BRC model $(\boldsymbol{\theta}'_m, \mathbf{m}')$ in which $p(y|\mathbf{x}, \boldsymbol{\theta}'_m, \mathbf{m}') = p(y|\mathbf{x}, \boldsymbol{\theta}_m, \mathbf{m})$. We say that $(\boldsymbol{\theta}'_m, \mathbf{m}')$ is a *Bayesian embedded regression/classification* (BERC) *model obtained from* $(\boldsymbol{\theta}_m, \mathbf{m})$.

Several researchers have suggested using BERC models, at least implicitly (see Bishop, 1995, Chapter 10, and references therein). An example of a BERC model obtained from a naive Bayes model is shown in Figure 2. If a BERC model $(\boldsymbol{\theta}'_m, \mathbf{m}')$ is obtained from a model which is itself a BERC model, we refer to $(\boldsymbol{\theta}'_m, \mathbf{m}')$ as a trivial BERC model. The BERC model in Figure 2 is non-trivial.

For any Bayesian network with finite-state variables, it is not difficult to obtain its corresponding BERC model. Let $X_1, \ldots, X_{n_h}, Y, X_{n_h+1}, \ldots, X_n$ be a total ordering on the

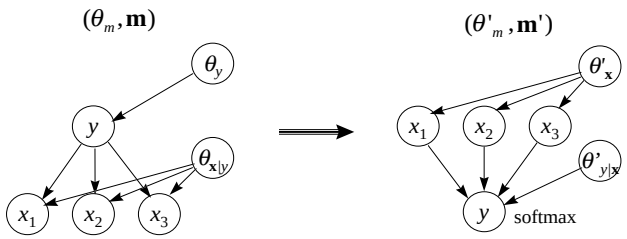


Figure 2: A BERC model obtained from a naive-Bayes model.

variables that is consistent with $\mathbf{m}$, such that Y appears as late as possible in the ordering. The latter condition says that the node corresponding to Y is an ancestor of each of the nodes corresponding to $X_{n_h+1}, \ldots, X_n$. Given this ordering, we can factor the joint distribution for $Y, X_1, \ldots, X_n$ as follows:

$$p(y, \mathbf{x}|\boldsymbol{\theta}_m, \mathbf{m}) = \left(\prod_{i=1}^{n_h} p(x_i|\mathbf{pa}_i, \boldsymbol{\theta}_m, \mathbf{m}) \right) \; p(y|\mathbf{pa}_y, \boldsymbol{\theta}_m, \mathbf{m}) \; \left(\prod_{i=n_h+1}^{n} p(x_i|\mathbf{pa}_i, \boldsymbol{\theta}_m, \mathbf{m}) \right)$$

where Y does not appear in any parent set $\mathbf{Pa}_i$ in the first product. Normalizing to obtain $p(y|\mathbf{x}, \boldsymbol{\theta}_m, \mathbf{m})$, taking a ratio, and canceling like terms, we obtain

$$\lambda_{k\mathbf{x}} = \log \frac{\theta(y^k|\mathbf{pa}_y)}{\theta(y^1|\mathbf{pa}_y)} + \sum_{i=n_h+1}^{n} \log \frac{\theta(x_i|\mathbf{pa}_i^k)}{\theta(x_i|\mathbf{pa}_i^1)} \tag{8}$$

where $\mathbf{pa}_i^k$ is a configuration of $\mathbf{Pa}_i$ in which $y = y^k$, $k = 1, \ldots, r$. (Depending on $\mathbf{m}$, some of the terms in the sum may cancel as well.) We can trivially rewrite Equation 8 as

$$\lambda_{k\mathbf{x}} = I(x_1) \cdots I(x_n) \log \frac{\theta(y^k|\mathbf{pa}_y)}{\theta(y^1|\mathbf{pa}_y)} + \sum_{i=n_h+1}^{n} \log \frac{\theta(x_i|\mathbf{pa}_i^k)}{\theta(x_i|\mathbf{pa}_i^1)} \tag{9}$$

Equation 9 shows that an BERC model is a *polynomial softmax regression on the indicator variables* $I(x_1), \ldots, I(x_n)$. Note that there are polynomial softmax regressions that cannot be obtained from any Bayesian network.

Although BERC models are convenient, we find non-trivial BERC models to be problematic. In particular, consider a BERC model $(\boldsymbol{\theta}'_m, \mathbf{m}')$ obtained from a non-BERC model $(\boldsymbol{\theta}_m, \mathbf{m})$. Whereas in the BERC model, observations of $\mathbf{X}$ are necessarily uninformative about $\boldsymbol{\theta}_{y|\mathbf{x}}$, such observations may be informative in the original model $(\boldsymbol{\theta}_m, \mathbf{m})$. Thus, in constructing the BERC model, we may be ignoring parts of our prior knowledge that are important for learning.

To illustrate this point, consider the naive-Bayes model for binary variables Y, X_1, X_2, X_3. The mapping from $\boldsymbol{\theta}_m$ to $\boldsymbol{\theta}_{\mathbf{x}}$ is given by

$$
\begin{aligned}
\theta(x_1^1, x_2^1, x_3^1) &= \theta(y^1)\theta(x_1^1|y^1)\theta(x_2^1|y^1)\theta(x_3^1|y^1) + \theta(y^2)\theta(x_1^1|y^2)\theta(x_2^1|y^2)\theta(x_3^1|y^2) \qquad (10)\\
\theta(x_1^1, x_2^1, x_3^2) &= \theta(y^1)\theta(x_1^1|y^1)\theta(x_2^1|y^1)(1-\theta(x_3^1|y^1)) + \theta(y^2)\theta(x_1^1|y^2)\theta(x_2^1|y^2)(1-\theta(x_3^1|y^2))\\
\theta(x_1^1, x_2^2, x_3^1) &= \theta(y^1)\theta(x_1^1|y^1)\theta(x_2^1|y^1)\theta(x_3^1|y^1) + \theta(y^2)\theta(x_1^1|y^2)\theta(x_2^1|y^2)\theta(x_3^1|y^2)\\
\theta(x_1^1, x_2^2, x_3^2) &= \theta(y^1)\theta(x_1^1|y^1)(1-\theta(x_2^1|y^1))(1-\theta(x_3^1|y^1)) + \theta(y^2)\theta(x_1^1|y^2)(1-\theta(x_2^1|y^2))(1-\theta(x_3^1|y^2))\\
\theta(x_1^2, x_2^1, x_3^1) &= \theta(y^1)(1-\theta(x_1^1|y^1))\theta(x_2^1|y^1)\theta(x_3^1|y^1) + \theta(y^2)(1-\theta(x_1^1|y^2))\theta(x_2^1|y^2)\theta(x_3^1|y^2)\\
\theta(x_1^2, x_2^1, x_3^2) &= \theta(y^1)(1-\theta(x_1^1|y^1))\theta(x_2^1|y^1)(1-\theta(x_3^1|y^1)) + \theta(y^2)(1-\theta(x_1^1|y^2))\theta(x_2^1|y^2)(1-\theta(x_3^1|y^2))\\
\theta(x_1^2, x_2^2, x_3^1) &= \theta(y^1)(1-\theta(x_1^1|y^1))\theta(x_2^1|y^1)\theta(x_3^1|y^1) + \theta(y^2)(1-\theta(x_1^1|y^2))\theta(x_2^1|y^2)\theta(x_3^1|y^2)
\end{aligned}
$$

where we have used $\theta(y), \theta(x_1, x_2, x_3)$, and $\theta(x_i|y)$ to denote $p(y|\boldsymbol{\theta}_m, \mathbf{m}), p(x_1, x_2, x_3|\boldsymbol{\theta}_m, \mathbf{m})$, and $p(x_i|y, \boldsymbol{\theta}_m, \mathbf{m})$, respectively.It is not difficult to show that the rank of the Jacobian matrix $\partial\boldsymbol{\theta}_{\mathbf{x}}/\partial\boldsymbol{\theta}_m$ is full (i.e., equal to the number of non-redundant parameters in $\boldsymbol{\theta}_m$) for almost all values of Θ_m (see, e.g., Geiger et al., 1996). It follows that, for almost every point $\boldsymbol{\theta}_m^*$ in $\boldsymbol{\theta}_m$, there is an inverse mapping from $\boldsymbol{\theta}_{\mathbf{x}}$ to $\boldsymbol{\theta}_m$ in a neighborhood around $\boldsymbol{\theta}_m^*$.[3] Consequently, the possible values that $\boldsymbol{\theta}_m$ (and hence $\boldsymbol{\theta}_{y|\mathbf{x}}$) can assume will depend on the value of $\boldsymbol{\theta}_{\mathbf{x}}$, and observations of $\mathbf{X}$ will inform $\boldsymbol{\theta}_{y|\mathbf{x}}$ through $\boldsymbol{\theta}_{\mathbf{x}}$.

In general, given two variables (random or otherwise) A and B, if the possible values that can be assumed by A depend on the value of B, then A is said to be *variationally dependent* on B. In our example, $\Theta_{y|\mathbf{x}}$ is variationally dependent on $\Theta_{\mathbf{x}}$. Such variational dependence is not limited to this example. For any model $(\boldsymbol{\theta}_m, \mathbf{m})$, if the rank of the Jacobian matrix for the mapping from $\boldsymbol{\theta}_m$ to $\boldsymbol{\theta}_{\mathbf{x}}$ is full, then Θ_m (and hence $\Theta_{y|\mathbf{x}}$) is variationally dependent on $\boldsymbol{\theta}_{\mathbf{x}}$. Geiger et al. (1996) conjecture that, for naive-Bayes models in which all variables are binary, the rank of the Jacobian matrix for the mapping from $\boldsymbol{\theta}_m$ to $\boldsymbol{\theta}_{\mathbf{x}}$ is almost everywhere full. In addition, Goodman (1974) and Geiger et al. (1996) could identify only one naive-Bayes model in which the Jacobian matrix was not of full rank almost everywhere. Thus, the use of non-trivial BERC models—at least those obtained from most naive Bayes models—is suspect.

Note that our remarks extend to non-Bayesian analyses. For example, in a classical analysis, a polynomial softmax regression should not be substituted for a Bayesian network. In the former model, $\Theta_{y|\mathbf{x}}$ and $\Theta_{\mathbf{x}}$ are variationally independent. In the latter model, $\Theta_{y|\mathbf{x}}$ and $\Theta_{\mathbf{x}}$ are variationally dependent, and observations of $\mathbf{X}$ often will influence the estimate of $\boldsymbol{\theta}_{y|\mathbf{x}}$. More generally, conditional models—often referred to as regression/classification models—should not be used without consideration of variational dependencies that may arise from the joint model.

[3] The parameters $\boldsymbol{\theta}_m$ are said to be *locally identifiable* given observations of $\mathbf{X}$ (e.g., Goodman, 1974).

4 Learning Regression/Classification Models: Averaging Versus Selection

Now that we have examined several classes of models for the regression/classification task, let us concentrate on Bayesian methods for learning such models.

First, consider *model averaging*. Given a random sample D from the true distribution of Y and $\mathbf{X}$, we compute the posterior distributions for each $\mathbf{m}$ and $\boldsymbol{\theta}_m$ using Bayes' rule:

$$p(\mathbf{m}|D) = \frac{p(\mathbf{m})\ p(D|\mathbf{m})}{\sum_{m'} p(\mathbf{m}')\ p(D|\mathbf{m}')}$$

$$p(\boldsymbol{\theta}_m|D,\mathbf{m}) = \frac{p(\boldsymbol{\theta}_m|\mathbf{m})\ p(D|\boldsymbol{\theta}_m,\mathbf{m})}{p(D|\mathbf{m})}$$

where

$$p(D|\mathbf{m}) = \int p(D|\boldsymbol{\theta}_m,\mathbf{m})\ p(\boldsymbol{\theta}_m|\mathbf{m})\ d\boldsymbol{\theta}_m$$

With these quantities in hand, we can determine the conditional distribution for Y given $\mathbf{X}$ in the next case to be seen by averaging over all possible model structures and their parameters:

$$p(y|\mathbf{x},D) = \sum_m p(\mathbf{m}|D)\ p(y|\mathbf{x},D,\mathbf{m}) \tag{11}$$

$$p(y|\mathbf{x},D,\mathbf{m}) = \int p(y|\mathbf{x},\boldsymbol{\theta}_m,\mathbf{m})\ p(\boldsymbol{\theta}_m|D,\mathbf{m})d\boldsymbol{\theta}_m \tag{12}$$

Note that joint analysis is handled in essentially the same way. For example, to determine the joint distribution of Y and $\mathbf{X}$ in the next case to be seen, we use

$$p(y,\mathbf{x}|D) = \sum_m p(\mathbf{m}|D)p(y,\mathbf{x}|D,\mathbf{m}) \tag{13}$$

$$p(y,\mathbf{x}|D,m) = \int p(y,\mathbf{x}|\boldsymbol{\theta}_m,\mathbf{m})\ p(\boldsymbol{\theta}_m|D,\mathbf{m})d\boldsymbol{\theta}_m \tag{14}$$

Model averaging, however, is not always appropriate for an analysis. For example, only one or a few models may be desired for domain understanding or for fast prediction. In these situations, we select one or a few "good" model structures from among all possible models, and use them as if they were exhaustive. This procedure is known as *model selection* when one model is chosen, and *selective model averaging* when more than one model is chosen. Of course, model selection and selective model averaging are also useful when it is impractical to average over all possible model structures.

When our goal is model selection, a "good" model for joint analysis may not be a good model for regression/classification, and vice versa. Scores that define "good" model structures are commonly known as *criteria*. A criterion commonly used for joint analysis

is the logarithm of the relative posterior probability of the model structure $\log p(\mathbf{m}, D) = \log p(\mathbf{m}) + \log p(D|\mathbf{m})$. This criterion is *global* in the sense that it is equally sensitive to possible dependencies among all variables. Criteria for regression/classification, should be *local* in the sense that they concentrate on how well $\mathbf{X}$ classifies Y. In the following section, we examine two such criteria.

5 Prequential Criteria for Regression/Classification

The criteria that we discuss can be understood in terms of Dawid's (1984) predictive sequential or *prequential* method. A simple example of this method, applied to joint analysis, yields the posterior-probability criterion. Let us consider this example first.

To simplify the discussion, let us assume that that $p(\mathbf{m})$ is uniform, so that the joint-analysis criterion reduces to the log-marginal-likelihood $\log p(D|\mathbf{m})$.[4] From the chain rule of probability, the log marginal likelihood is given by

$$\log p(D|\mathbf{m}) = \sum_{l=1}^{N} \log p(y_l, \mathbf{x}_l | y_1, \mathbf{x}_1, \ldots, y_{l-1}, \mathbf{x}_{l-1}, \mathbf{m})$$

The term $p(y_l, \mathbf{x}_l | y_1, \mathbf{x}_1, \ldots, y_{l-1}, \mathbf{x}_{l-1}, \mathbf{m})$ is the prediction for $(y_l, \mathbf{x}_l)$ made by model structure $\mathbf{m}$ after averaging over its parameters (Equation 14). The log of this term can be thought of as the utility for this prediction.[5] Thus, a model structure with the highest log marginal likelihood is also a model structure that is the best sequential predictor of the data D given the logarithmic utility function.

Let us now consider local criteria that are more appropriate for the task of regression/classification. To keep the discussion brief, we discuss only the logarithmic utility function, although other utility functions may be more reasonable for a given problem. At least two prequential criteria are reasonable. In one situation, we imagine that we see pairs $(y_l, \mathbf{x}_l)$ sequentially. As a result, we obtain a criterion that Spiegelhalter et al. (1993) call a *conditional node monitor*:

$$\mathrm{CNM}(D, \mathbf{m}) = \sum_{l=1}^{N} \log p(y_l | \mathbf{x}_l, y_1, \mathbf{x}_1, \ldots, y_{l-1}, \mathbf{x}_{l-1}, \mathbf{m}) \tag{15}$$

In another situation, we imagine that we first see all of the input data $\mathbf{x}_1, \ldots, \mathbf{x}_N$, and then see the class data sequentially. Consequently, we obtain the following *class sequential*

[4] The generalization to non-uniform model priors is straightforward.

[5] The utility $\log x$ is also known as a *scoring rule*. Bernardo (1979) shows that this scoring rule has several desirable properties.

criterion:

$$\mathrm{CSC}(D, \mathbf{m}) = \sum_{l=1}^{N} \log p(y_l | y_1, \ldots, y_{l-1}, \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{m}) \tag{16}$$

Buntine (1993) used this criterion for selection among decision-tree structures.

Spiegelhalter et al. (1993) describe a set of assumptions—essentially, parameter independence and Dirichlet priors—under with the conditional node monitor can be computed efficiently in closed form. Under these same assumptions, the exact computation of the class sequential criterion is exponential in the sample size N. Monte-Carlo or asymptotic techniques can be used to perform the computation for large N (see, e.g., Heckerman, 1995).

We have applied both criteria to small Bayesian networks and small data sets chosen arbitrarily. In all cases, we have found that the two criteria differ. Nonetheless, there are conditions under which the two criteria are the same. In particular, we can rewrite the two criteria as follows:

$$\mathrm{CNM}(D, \mathbf{m}) = \sum_{l=1}^{N} \log \frac{p(y_l, \mathbf{x}_l | y_1, \mathbf{x}_1, \ldots, y_{l-1}, \mathbf{x}_{l-1}, \mathbf{m})}{p(\mathbf{x}_l | y_1, \mathbf{x}_1, \ldots, y_{l-1}, \mathbf{x}_{l-1}, \mathbf{m})} \tag{17}$$

$$\mathrm{CSC}(D, \mathbf{m}) = \log \frac{p(y_1, \ldots, y_N, \mathbf{x}_1, \ldots, \mathbf{x}_N | \mathbf{m})}{p(\mathbf{x}_1, \ldots, \mathbf{x}_N | \mathbf{m})} = \sum_{l=1}^{N} \log \frac{p(y_l, \mathbf{x}_l | y_1, \mathbf{x}_1, \ldots, y_{l-1}, \mathbf{x}_{l-1}, \mathbf{m})}{p(\mathbf{x}_l | \mathbf{x}_1, \ldots, \mathbf{x}_{l-1}, \mathbf{m})} \tag{18}$$

Therefore, the two criteria will agree when

$$p(\mathbf{x}_l | y_1, \mathbf{x}_1, \ldots, y_{l-1}, \mathbf{x}_{l-1}, \mathbf{m}) = p(\mathbf{x}_l | \mathbf{x}_1, \ldots, \mathbf{x}_{l-1}, \mathbf{m}) \tag{19}$$

for $l = 0, \ldots, N-1$. It is not difficult to show that Equation 19 holds whenever $(\boldsymbol{\theta}_m, \mathbf{m})$ is a BRC model. Thus, the two criteria agree for BRC models.

6 Discussion

Several researchers have demonstrated that Bayesian networks for both the joint analysis and regression/classification tasks provide better predictions when local distribution functions are encoded with a small number of parameters, as is the case with the use of decision trees, decision graphs, and causal-independence models (e.g., Friedman and Goldszmidt, 1996; Chickering et al., 1997; Meek and Heckerman, 1997). Despite our theoretical objections to the use of BERC models, they offer another parsimonious parameterization of local distribution functions, and may lead to better predictions in practice. For example, polynomial softmax regressions may be useful when a node and its parents are discrete. Experiments are needed to investigate these possibilities.

Acknowledgments

We thank Max Chickering for useful discussions.

References

Bernardo, J. (1979). Expected information as expected utility. *Annals of Statistics*, 7:686–690.

Bishop, C. (1995). *Neural networks for pattern recognition*. Clarendon Press, Oxford.

Buntine, W. (1993). Learning classification trees. In *Artificial Intelligence Frontiers in Statistics: AI and statistics III*. Chapman and Hall, New York.

Chickering, D., Heckerman, D., and Meek, C. (1997). A Bayesian approach to learning Bayesian networks with local structure. In *Proceedings of Thirteenth Conference on Uncertainty in Artificial Intelligence,* Providence, RI. Morgan Kaufmann.

Dawid, P. (1984). Present position and potential developments: some personal views. Statistical theory. The prequential approach (with Discussion). *Journal of the Royal Statistical Society A*, 147:178–292.

Friedman, N. and Goldszmidt, M. (1996). Building classifiers using Bayesian networks. In *Proceedings AAAI-96 Thirteenth National Conference on Artificial Intelligence,* Portland, OR, pages 1277–1284. AAAI Press, Menlo Park, CA.

Geiger, D., Heckerman, D., and Meek, C. (1996). Asymptotic model selection for directed networks with hidden variables. In *Proceedings of Twelth Conference on Uncertainty in Artificial Intelligence,* Portland, OR. Morgan Kaufmann.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (1995). *Bayesian Data Analysis*. Chapman and Hall.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61:215–231.

Heckerman, D. (1995). A tutorial on learning Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, Redmond, WA. Revised January, 1996.

Meek, C. and Heckerman, D. (1997). Structure and parameter learning for causal independence and causal interaction models. In *Proceedings of Thirteenth Conference on Uncertainty in Artificial Intelligence,* Providence, RI. Morgan Kaufmann.

Spiegelhalter, D., Dawid, A., Lauritzen, S., and Cowell, R. (1993). Bayesian analysis in expert systems. *Statistical Science*, 8:219–282.