CS281A/Stat241A: Statistical Learning Theory

Density Estimation, Regression and Classification (09/28/04)

Lecturer: Michael I. Jordan *Scribe: Karthik Lakshminarayanan*

In this class, we look at density estimation for discrete data, taking the multinomial distribution as an example. We then consider mixture models, and look at density estimation when we model the data using mixture models. Following density estimation, we examine regression, i.e., how to model the dependence of the output variable on an input (or a covariate) variable. Finally, we introduce the basics of classification.

1 Density Estimation

1.1 Bayesian Density Estimation for Discrete Data

In this section, taking the multinomial distribution as an example, we demonstrate a Bayesian approach to density estimation. We use a parameterized prior distribution as before and compute the corresponding posterior. Earlier, both the prior as well as the posterior had the same distribution—the Gaussian distribution. In this case, to achieve the same property we need to pick a prior distribution so that when multiplied with a multinomial, the result is the same prior distribution. This property of a prior distribution for a particular likelihood distribution is called *conjugacy*. We observe that the Dirichlet distribution, $p(\theta) = C(\alpha)\theta_1^{\alpha_1 - 1} \ldots \theta_M^{\alpha_M - 1}$, where $\alpha = (\alpha_1, \ldots, \alpha_M)$ is the hyper-parameter, would achieve this desirable property.

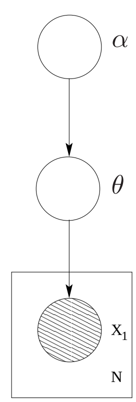

Figure 1: Graphical model for multinomial density estimation with Dirichlet prior with hyper-parameter α.

Now, the posterior probability can be calculated as follows:

$$\begin{aligned}
p(\theta | x_1, \ldots, x_N, \alpha) &\propto \Big(\prod_{i=1}^{N} \theta_1^{x_i^1} \ldots \theta_k^{x_i^k}\Big)\Big(\prod_{j=1}^{k} \theta_j^{\alpha_j - 1}\Big) \\
&= \prod_{j=1}^{k} \theta_j^{\left(\sum_i x_i^j + \alpha_j - 1\right)}
\end{aligned}$$

The above distribution is a Dirichlet distribution with parameters $\left(\sum_{i=1}^{N} x_i^j + \alpha_j\right)$. We can also compute the appropriate normalization factor. Observe that the posterior is obtained by simply adding $\sum_{i=1}^{N} x_i^j$ to parameter α_j within the Dirichlet distribution.

If we were to compute the MLE for the distribution, we would have:

$$\hat{\theta}_{k,ML} = \frac{\sum_{i=1}^{N} x_i^k}{N}$$

Observe that if the domain of the model was word distributions within text documents, using the MLE, a new document will have probability zero when there is some new word in it (which is a highly probable event). Using the MLE is perhaps not a great idea. Instead, we compute the Bayesian estimate after a prior belief in the form of α is added.

$$\begin{aligned}
\hat{\theta}_{k,Bayes} &= \frac{\sum_{i=1}^{N} x_i^k + \alpha_k}{\sum_k \sum_{i=1}^{N} x_i^k + \alpha_k} \\
&= \frac{\sum_i x_i^k + \alpha_k}{\sum_k \sum_i x_i^k + \sum_k \alpha_k} \\
&= \frac{N}{N+|\alpha|}\bar{x}_k + \frac{|\alpha|}{N+|\alpha|}\frac{\alpha_k}{|\alpha|} \\
&= \beta \bar{x}_k + (1-\beta)\frac{\alpha_k}{|\alpha|} \\
&= \beta\ [\text{MLE}] + (1-\beta)\ [\text{prior belief}]
\end{aligned}$$

Adding a prior belief in the form of α produces an estimate that is a weighted mean of prior belief and the maximum likelihood estimate. As as aside, we note that this phenomenon is more general; it happens with all members of the exponential family (of which multinomial and Dirichlet distributions are members).

1.2 Mixture Models

Thus far, we have considered the Gaussian and multinomial distributions to model data. Clearly, they are not be the best choices in all cases. Later, we discuss maximum likelihood and Bayesian estimates for the exponential family, a general family that includes the the Gaussian and multinomial distributions.

However, even this larger family may be inadequate in some cases. For instance, one might want to model data that is bimodal. We choose a prior that is a convex sum of different distributions.

$$p(\theta) = \sum_{l=1}^{L} \pi_l dir(\theta|\alpha_l), \quad 0 < \pi_l, \sum \pi_l = 1$$

Note that computing the posterior is similar to previous case — using a mixture of Dirichlet distributions as prior gives a mixture of Dirichlet as posterior.

$$\begin{aligned}
P(\theta|x_1,\ldots,x_N) &\propto \Big(\prod_i \prod_j \theta_j^{x_i^j}\Big)\Big(\sum_l \pi_l dir(\theta|\alpha_l)\Big) \\
&= \sum_l \pi_l \Big[\big(\prod_i \prod_j \theta_j^{x_i^j}\big) dir(\theta|\alpha_l)\Big] \\
&= \sum_l \pi_l dir(\theta|\alpha_l + \sum_{n=1}^{N} x_n^l)
\end{aligned}$$

2 Regression

Regression deals with how to model the dependence of the output variable on an input (or a covariate) variable. We assume that we are given a set of observed data $\{(x_i, y_i)\}_{i=1}^N$, where x_i is an observation of an input variable and y_i is the corresponding output; we assume IID sampling for simplicity. Figure 2 shows the graphical representation.

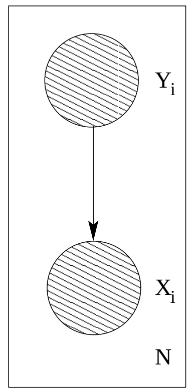

Figure 2: IID graphical model for regression.

The goal of regression is that we can predict y_{N+1} from x_{N+1}. In the frequentist approach, an estimate $\hat{\theta}$ is computed and prediction is done as $p(y_{N+1}|x_{N+1}, \hat{\theta})$. In the Bayesian approach, the posterior $p(\theta|x_1, y_1, \ldots, x_N, y_N)$ is computed. The prediction $p(y_{N+1}|x_{N+1})$ is computed as:

$$p(y_{N+1}|x_{N+1}) = \int p(y_{N+1}|x_{N+1}, \theta)\ p(\theta|x_1, y_1, \ldots, x_N, y_N)\ d\theta$$

2.1 Linear Regression

A linear regression model expresses y_n as a sum of a deterministic component (that depends on x_n) and a random component (that is independent of x_n) distributed according to a Gaussian distribution with mean 0 and variance σ^2; see Figure 3.

$$y_n = \beta^T x_n + \epsilon_n$$

In the linear regression case, we have:

$$\begin{aligned} p(y_i|x_i, \theta) &= \mathcal{N}(\theta^T x_i, \sigma^2) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\Big\{ -\frac{1}{2\sigma^2}(y_i - \beta^T x_i)^2 \Big\} \end{aligned}$$

Estimating the parameters for the regression problem is no different from the corresponding density estimation problem; we form the log likelihood, take derivative and set it to zero.

If $X = (x_1^T \ldots x_N^T)^T$ and the design matrix $Y = (y_1 \ldots y_N)^T$, the maximum likelihood estimate can be obtained by solving the normal equation:

$$X^T X \hat{\theta}_{ML} = X^T Y$$

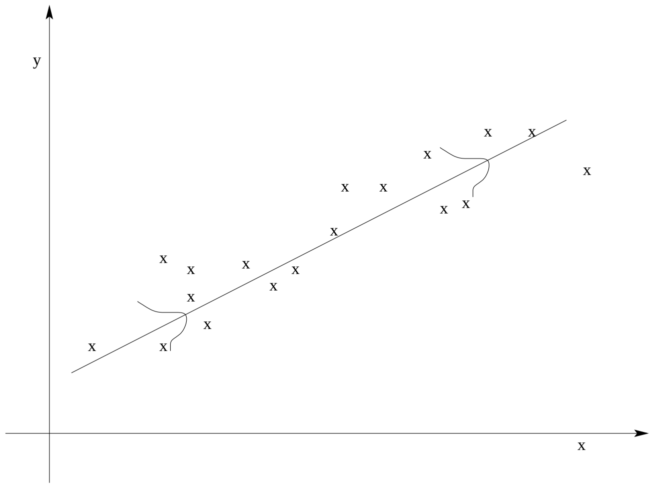

Figure 3: Linear regression model: represents conditional mean by the line as well as the input-independent random variation around the mean.

Note that the linearity we speak about is in the *parameters*, not the input variables. If the input X is non-linear, we can perform an appropriate fixed transformation from X to Z, adding the required parameters so that we can write the equations as $Y = \theta^T Z + \epsilon$.

3 Classification

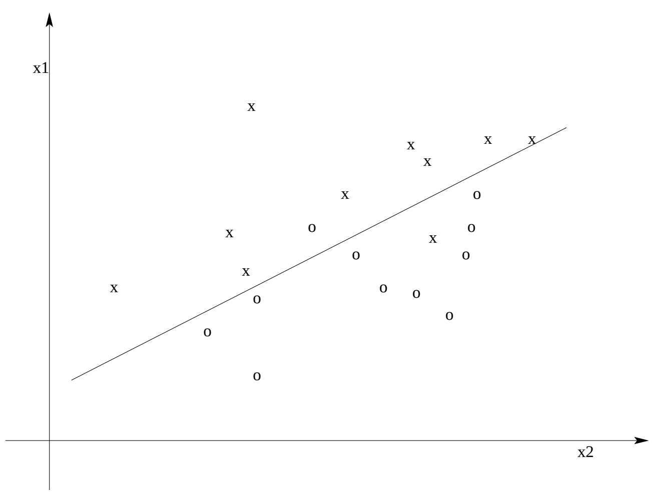

Figure 4: Classification: An example.

In classification, the response variable ranges over a finite set: as the name indicates we wish to classify the objects into one of the categories of responses. We refer to the covariate X as a *feature vector* and the discrete response Y as a *class label*.

There are two basic approaches to classification—generative and discriminative—as shown in Figure 5. In the generative approach, there is an arrow from from random variable Y to X; this approach is related to density

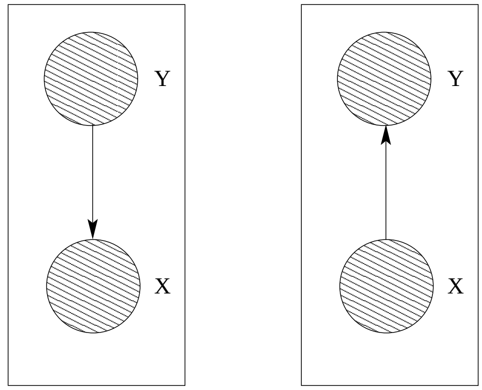

Figure 5: Graphical models for classification: (a) generative approach, and (b) discriminative approach.

estimation. For each discrete value of Y, we have a density $p(x|y)$ which is referred to as class-conditional density. The marginal $p(x)$, referred to as the prior, is required to compute the posterior, $p(y|x)$, of class Y. In the discriminative approach, there is an arrow from X to Y; this approach is related to regression. To classify an input, we can obtain the probability distribution of that input falling in each of the output class $(p(y|x))$ directly.

We now use the specific example of two classes, and Gaussian class-conditional densities with equal covariance matrices for the two classes. As we shall see later, the posterior $p(Y_i = 1|x_i, \theta)$ is a logistic function $1/(1 + e^{-\theta^T x_i})$. We shall study classification in greater detail in the next class.