CS281A/Stat241A: Statistical Learning Theory

Discrete Data (Lecture 10, September 30, 2002)

Lecturer: Michael I. Jordan *Scribes: Xiaotian Sun and Yun Zhou*

1 Multinomial/Dirichlet Distribution

Given data

$$x = (x_1, x_2, \ldots, x_N), \quad x_n^i \in \{0,1\}, \forall n = 1, 2, \ldots, N, \text{and } i = 1, 2, \ldots, M,$$

the *multinomial* distribution is given by

$$p(x|\theta) = \theta_1^{\sum_{n=1}^N x_n^1} \theta_2^{\sum_{n=1}^N x_n^2} \cdots \theta_M^{\sum_{n=1}^N x_n^M}. \tag{1}$$

The conjugate prior distribution for θ is the Dirichlet distribution:

$$p(\theta|\alpha) = C(\alpha)\theta_1^{\alpha_1 - 1}\theta_2^{\alpha_2 - 1} \cdots \theta_M^{\alpha_M - 1}, \tag{2}$$

where the normalization factor

$$C(\alpha) = \frac{\Gamma\left(\sum_{i=1}^M \alpha_1\right)}{\prod_{i=1}^M \Gamma(\alpha_i)}. \tag{3}$$

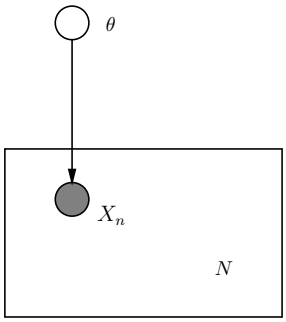

Figure 1: Estimator—Discrete data, multinomial distribution.

Posterior distribution

$$\begin{aligned} p(\theta|x) &\propto p(x|\theta)p(\theta|\alpha) \\ &= \theta_1^{\sum_{n=1}^N x_n^1 + \alpha_1 - 1} \theta_2^{\sum_{n=1}^N x_n^2 + \alpha_2 - 1} \cdots \theta_M^{\sum_{n=1}^N x_n^M + \alpha_M - 1}. \end{aligned} \tag{4}$$

Mean of $\mathrm{Di}(\alpha)$

$$\mathrm{E}[\theta_k] = \frac{\alpha_k}{\sum_{i=1}^M \alpha_i} \triangleq \frac{\alpha_k}{|\alpha|} \tag{5}$$

Covariance

$$\begin{aligned}\mathrm{Cov}(\theta_j, \theta_k) &\triangleq \mathrm{E}\big[\big(\theta_j - \mathrm{E}[\theta_j]\big)\big(\theta_k - \mathrm{E}[\theta_k]\big)\big] \\ &= \mathrm{E}[\theta_j\theta_k] - \mathrm{E}[\theta_j]\,\mathrm{E}[\theta_k] \\ &= \cdots .\end{aligned} \tag{6}$$

The maximum likelihood estimator

$$\hat{\theta}_{\mathrm{ML},k} = \frac{1}{N}\sum_{n=1}^{N} x_n^k. \tag{7}$$

The Bayesian estimator

$$\begin{aligned}\hat{\theta}_{\mathrm{Bayes},k} &= \frac{\sum_{n=1}^{N} x_n^k + \alpha_k}{\sum_{k=1}^{M}\left(\sum_{n=1}^{N} x_n^k + \alpha_k\right)} = \frac{\sum_{n=1}^{N} x_n^k + \alpha_k}{N + |\alpha|} \\ &= \frac{N/|\alpha|}{N/|\alpha| + 1}\hat{\theta}_{\mathrm{ML},k} + \frac{1}{N/|\alpha| + 1}\frac{\alpha_k}{|\alpha|}.\end{aligned} \tag{8}$$

The maximum *a posteriori* estimator

$$\hat{\theta}_{\mathrm{MAP},k} = \cdots . \tag{9}$$

Prediction

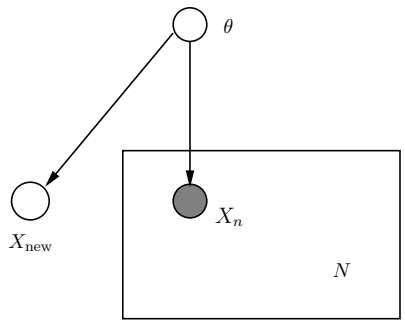

Figure 2: Predictor—Discrete data, multinomial distribution.

2 Poisson/Gamma Distribution

$$X_n \in \{0, 1, 2, \dots\}.$$

$X \sim \mathrm{Poisson}(\lambda)$, *i.e.*,

$$p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}. \tag{10}$$

Conjugate prior

$$p(\lambda|\alpha, \beta) = K(\alpha, \beta)\lambda^{\alpha-1}e^{-\beta\lambda}, \tag{11}$$

which is Gamma distribution.

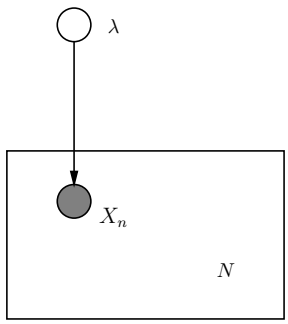

Figure 3: Estimator—Discrete data, Gamma distribution.

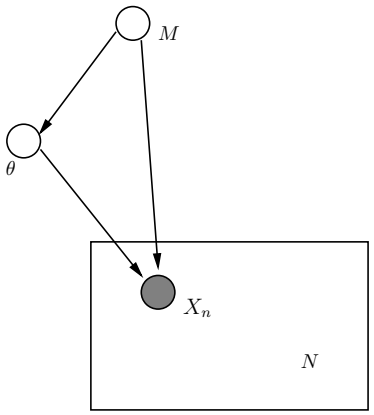

Figure 4: Model selection.

3 Model Selection

In Figure 4, M is the models, θ is the parameters, and x is the data.

A special case, where the data x is conditionally independent to the model M, given the parameters θ, as shown in Figure 5.

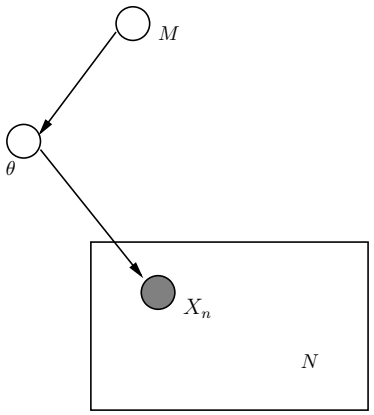

Figure 5: Model selection.

$$\begin{aligned} p(m|x) &= \int p(m,\theta|x)\ \mathrm{d}\theta \\ &= \int \frac{p(x|m,\theta)p(m,\theta)}{p(x)}\ \mathrm{d}\theta \\ &= \int \frac{p(x|\theta)}{p(x)} p(\theta|m)p(m)\ \mathrm{d}\theta. \end{aligned} \tag{12}$$

This implies that

$$\frac{p(m|x)p(x)}{p(m)} = \int p(x|\theta)p(\theta|m)\ \mathrm{d}\theta = p(x|m), \tag{13}$$

which is called the marginal likelihood. And the model estimator is given by

$$\hat{M} = \arg\max_m p(x|m). \tag{14}$$

Prediction with model averaging.

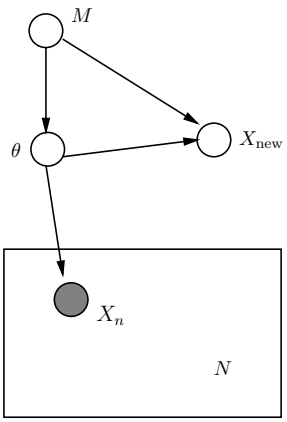

Figure 6: Prediction with model averaging.

$$\begin{aligned} p(x_{\text{new}}|x) &= \iint p(x_{\text{new}},\theta,m|x)\ \mathrm{d}\theta\,\mathrm{d}m \\ &= \iint \underbrace{p(x_{\text{new}}|\theta,m)}_{\text{Vote}} \underbrace{p(\theta|x,m)}_{\substack{\text{Parameter}\\ \text{posterior}}} \underbrace{p(m|x)}_{\substack{\text{Model}\\ \text{posterior}}}\ \mathrm{d}\theta\,\mathrm{d}m. \end{aligned} \tag{15}$$

4 Regression Model

Linear regression.

$$y_n = \beta^T x_n + \epsilon_n, \tag{16}$$

where $\epsilon_n \sim N(0,\sigma^2)$ is the noise and ϵ_n's are i.i.d. Our parameters

$$\theta = (\beta,\sigma^2). \tag{17}$$

Y_n is also Gaussian and

$$Y_n \sim N(\beta^T x_n, \sigma^2). \tag{18}$$

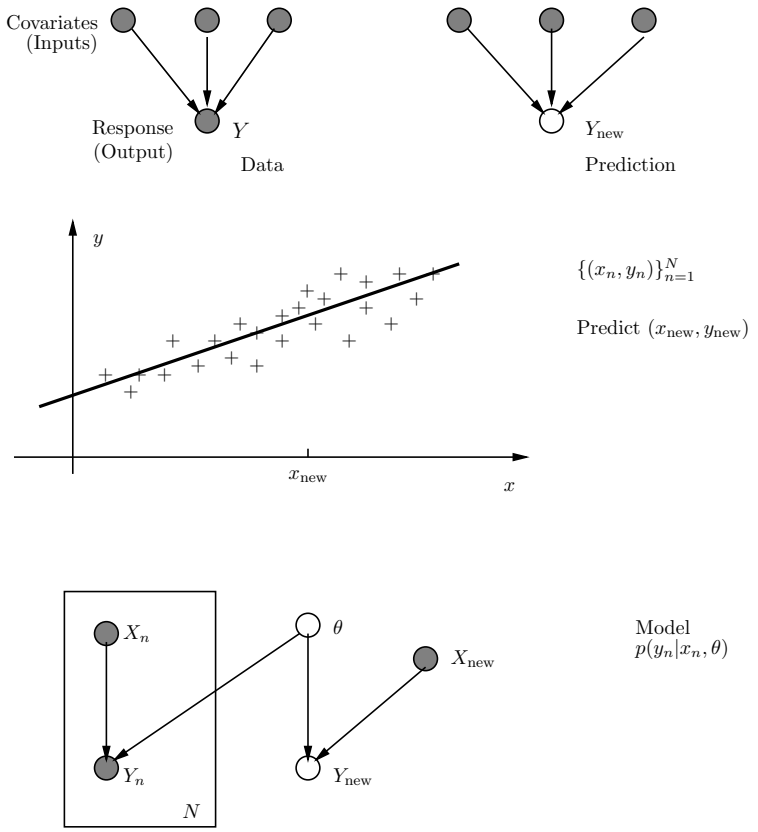

Figure 7: Regression model.

The model

$$p(y_n|x_n, \theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_n - \beta^T x_n)^2}. \tag{19}$$

Let

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \qquad X = \begin{pmatrix} \text{---} & x_1^T & \text{---} \\ \text{---} & x_2^T & \text{---} \\ & \vdots & \\ \text{---} & x_N^T & \text{---} \end{pmatrix}.$$

$$\hat{\theta}_{\text{ML}} = (X^T X)^{-1} X^T Y, \tag{20}$$

which is called the normal equation. This normal equation can be obtained as follows:

$$L(\beta) = \prod_{n=1}^{N} e^{-\frac{1}{2}(y_n - \beta^T x_n)^2}. \tag{21}$$

$$l(\beta) = \log L(\beta) = -\frac{1}{2} \sum_{n=1}^{N} (y_n - \beta^T x_n)^2. \tag{22}$$

Set $\frac{\partial l}{\partial \beta} = 0 \Longrightarrow$ normal equation.