CS281A/Stat241A: Statistical Learning Theory

Parameter Estimation (9/21/04)

Lecturer: Michael I. Jordan *Scribes: Xingyuan Chen and Dengfeng Sun*

1 Parameter estimation

1.1 Model M, Parameter θ, and Observed Data X

The relationship between (M, θ) and X is shown in Figure 1.

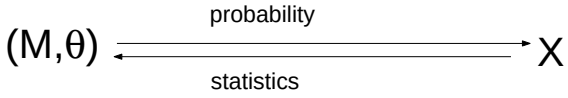

Figure 1: Relationship between (M, θ) and X.

For example, M could be a grpahical model, θ could be some numerical entities and X could be some observed data set.

1.2 Bayes' Rule

If we think of the parameter θ as a random variable, then we can write out Bayes' rule in terms of the posterior probability of θ ($p(\theta|x)$), the likelihood ($p(x|\theta)$), and the prior probability of parameter θ ($p(\theta)$):

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \tag{1}$$

Bayes' rule defines a relationship between the three components:

$$posterior \propto likelihood * prior \tag{2}$$

2 Prediction Methods

2.1 Bayesian Method

Given data set $\mathcal{X} = \{X_1, \cdots, X_N\}$, we want to predict the next $X : X_{N+1}$.

Let X_i's be (conditionally) independent and identically distributed given θ, as illustrated in Figure 2. Since X_i's are i.i.d., $p(x_i|\theta)$ is same for all i.

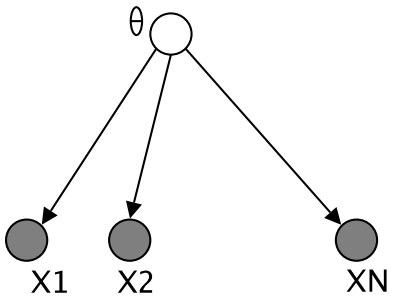

Figure 2: X_i i.i.d. given θ.

2.2 Exchangeability

X_i's are exchangeable means

$$p(x_1, \cdots, x_N) = p(x_{\pi_1}, \cdots, x_{\pi_N}), \tag{3}$$

Notice that π_i here does not mean the parents; it means an arbitrary permutation.

Remark: *exchangeable* is weaker than i.i.d..

de Finetti claims: **exchangeability** $\Longleftrightarrow p(x_1, \cdots, x_N) = \int d\theta p(\theta) \prod_{i=1}^{N} p(x_i|\theta)$ **for some** $p(\theta)$.

Let us denote Figure 2 using plate notation, as in Figure 3 where N is the number of X_i's.

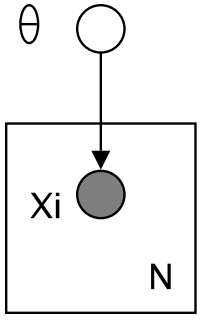

Figure 3: X_i i.i.d. given θ.

The *prediction* problem is shown in Figure 4.

In Figure 4, the posterior probability is as follows:

$$\begin{aligned}
p(x_{N+1}|x_1, \cdots, x_N) &= \int p(x_{N+1}, \theta|x_1, \cdots, x_N) d\theta \\
&= \int p(x_{N+1}|\theta, x_1, \cdots, x_N) p(\theta|x_1, \cdots, x_N) d\theta \\
&\text{(recall: } p(a, b|c) = p(a|b, c)p(b|c); p(a, b) = p(a|b)p(b)\text{)} \\
&= \int \underbrace{p(x_{N+1}|\theta)}_{\text{using conditional independence}} \underbrace{p(\theta|x_1, \cdots, x_N)}_{\text{posterior}} d\theta.
\end{aligned}$$

It can be difficult to do the above integral.

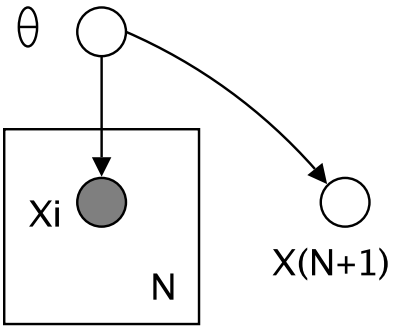

Figure 4: Prediction problem illustration.

3 Model Averaging/Model Selection

$$\left.\begin{array}{c} p(x|\theta, M) \\ p(\theta|M) \\ p(M) \end{array}\right\} p(x, \theta, M)$$

Relations of M, θ and X are shown in Figure 5.

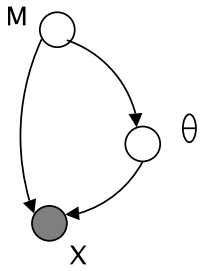

Figure 5: Graphical Representation.

Marginal likelihood:

$$p(x|M) = \int p(x, \theta|M) d\theta = \int d\theta p(x|\theta, M) p(\theta|M)$$

Posterior over models:

$$p(M|x) = \frac{p(x|M)p(M)}{p(x)}$$

Prediction:

$$p(x_{N+1}|x_1, \cdots, x_N) = \int \int p(x_{N+1}|\theta, M) p(\theta|x_1, \cdots, x_N, M) p(M|x_1, \cdots, x_N) d\theta dM$$

3.1 Frequentist Philosophy

Come back to $\theta \longrightarrow X$. Given X, define an estimator $\hat{\theta}(X)$ and show that it performs well on average (average over all possible data X)[1].

[1] This is one aspect that distinguishes frequentist from Bayesian.

3.1.1 Loss

A loss function defines the error associated with $\hat{\theta}(X)$, the parameter estimated from data X, versus the "real" θ (in the frequentist philosophy, the θ that actually generated the data). It is written $l(\theta, \hat{\theta}(X))$.

Squared error is an example of a loss function:

$$l(\theta, \hat{\theta}(X)) = (\theta - \hat{\theta}(X))^2. \tag{4}$$

3.1.2 Frequentist Risk

Frequentist risk is the expected loss for the parameter estimated from X, $\hat{\theta}(X)$, versus the "real" parameter θ, averaged over all possible data X.

$$\begin{aligned} R(\theta) &= E_\theta l(\theta, \hat{\theta}(X)) \\ &:= \int l(\theta, \hat{\theta}) p(x|\theta) dx \end{aligned}$$

Figure 6 shows an example of the risk function.

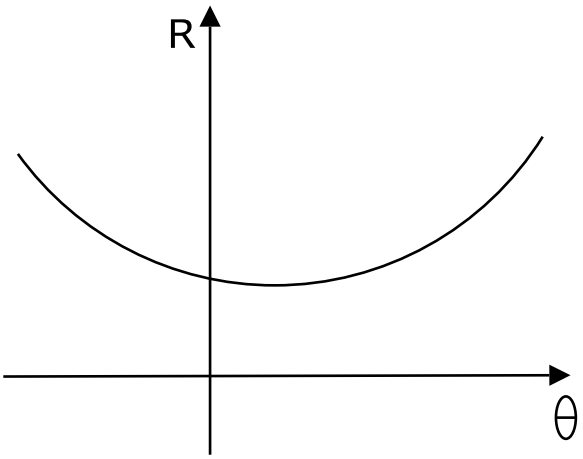

Figure 6: A risk function.

In Figure 7, comparison between two methods with corresponding risk functions $R_1(\theta)$, $R_2(\theta)$ and estimators $\hat{\theta}_1(X)$, $\hat{\theta}_2(X)$ is shown. It can be easily seen that method 2 is better than method 1 since $R_1(\theta)$ is greater than $R_2(\theta)$ for all θ. Usually curves cross and we could not simply say which method is better.

Concept: *Unbiased* means $E_\theta \hat{\theta}(X) = \theta$.

Two examples of frequentist parameter estimators are maximum likelihood estimation and minimax estimation.

Using the concept of risk, we can also define a Bayesian version of loss:

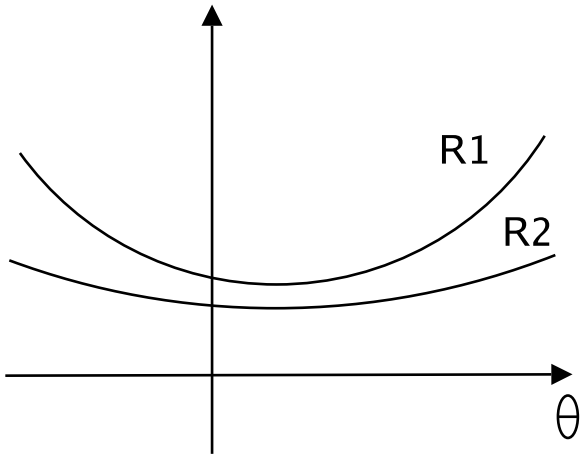

Figure 7: The risk of two estimation methods.

$$
\begin{aligned}
r &= \int R(\theta) \quad \underbrace{p(\theta)}_{\text{“drops off”}} \quad d\theta \\
&= \int \int l(\theta, \hat{\theta}) p(x|\theta) p(\theta) dx d\theta \\
&= \int dx \int \underbrace{d\theta l(\theta, \hat{\theta}) p(x|\theta) p(\theta)}_{\text{Bayes average loss}}
\end{aligned}
$$

Notice we average over the parameter values, not the data.

3.2 Frequentist vs. Bayesian

Frequentist: consider all possible data, and take the average of the estimator over the possible data; estimated values are given as specific values.

Bayesian: consider data fixed. Estimated parameters are characterized by a posterior distribution over possible parameter values.

4 Examples of Estimators

4.1 Maximum Likelihood Estimator

$$\hat{\theta}_{\text{ML}} = \arg\max_{\theta} p(x|\theta) \tag{5}$$

4.2 MAP Estimator

$$\begin{aligned}\hat{\theta}_{\text{MAP}} &= \arg\max_{\theta}[p(x|\theta)p(\theta)] \\ &= \arg\max_{\theta} p(\theta|x)\end{aligned} \tag{6}$$

Remark:

$$\arg\max_{\theta}[p(x|\theta)p(\theta)] = \arg\max_{\theta} \frac{p(x|\theta)p(\theta)}{p(x)} \tag{7}$$

since $p(x)$ is constant.

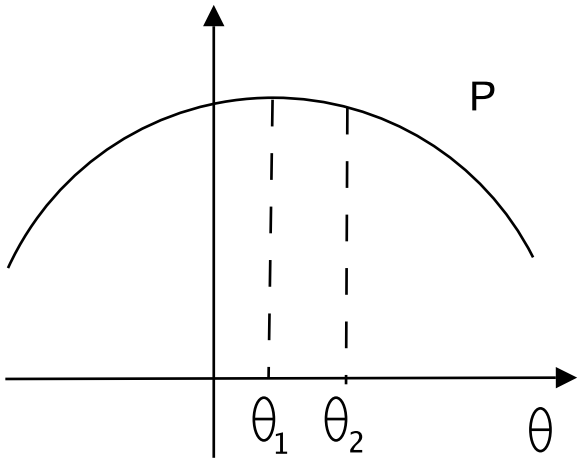

Figure 8: MAP and Bayes estimators ($\theta_1 = \hat{\theta}_{\text{MAP}}$, $\theta_2 = \hat{\theta}_{\text{Bayes}}$).

4.3 Bayes Estimator

$$\hat{\theta}_{\text{Bayes}}(X) = \int \theta p(\theta|x) d\theta \tag{8}$$

4.4 Various Robust Estimators

Trimmed means estimator: eliminate top and bottom 10% of x and estimate mean from the middle 80%.

4.5 Example — Gaussian Mean

$$p(x_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

$$p(x|\mu, \sigma^2) = \prod_{i=1}^{N} p(x_i|\mu, \sigma^2)$$

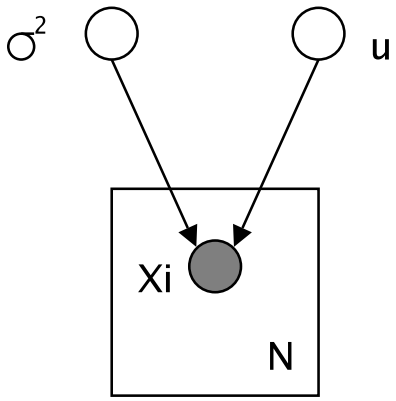

Figure 9: Gaussian Mean.

It can be proved (exercise) that maximum likelihood estimator of μ is:

$$\arg\max_{\mu} p(x|\mu,\sigma^2) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

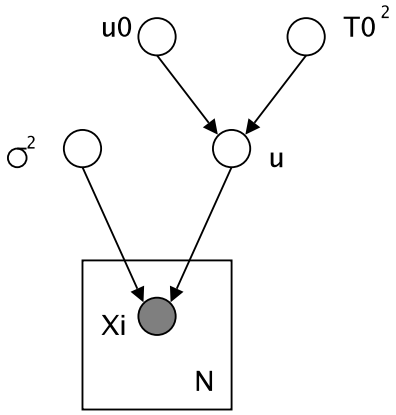

Figure 10: Gaussian distributed X, with a Gaussian prior on the mean μ.

If we further assume that μ follows a normal distribution as shown in Figure 10, then

$$\begin{aligned} p(\mu|\mu_0,\tau_0) &= \frac{1}{\sqrt{2\pi}\tau_0} e^{-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2} \\ p(\mu|x_1,\cdots,x_N) &\sim \mathcal{N}(f(\mu_0+\bar{x}),\tau_0)) \end{aligned} \tag{9}$$

where we will derive the function f during the next class.