Stat210A: Theoretical Statistics **Lecture Date: September 2, 2014**

Lecture 2: Risks and Exponential Families

Lecturer: Michael I. Jordan *Scribe: Johnny Hong*

1 Risks

Let $L(\theta, \delta(X))$ be a loss function. Recall the risk function:

$$R(\theta, \delta) = E_\theta L(\theta, \delta(X)) \tag{1}$$

This is called the *frequentist risk* because we consider θ to be fixed(unknown) and take expectation of the data. In contrast, for *Bayesian risks*, we take expectation *conditioning on data*.

Introduce the weighting function $\pi(\theta)$. Define the *Bayes risk* as

$$R(\delta, \pi) = \int R(\theta, \delta)\pi(\theta)d\theta \tag{2}$$

Recall from Bayes' Theorem, we have

$$p(\theta|x) \propto p(x|\theta)\pi(\theta) \tag{3}$$

where $p(\theta|x)$ is called the posterior, $p(x|\theta)$ is the likelihood function, and $\pi(\theta)$ is the prior.

Define the *Bayesian/Posterior risk*[1] as

$$r(x, \pi) = \int L(\theta, \delta(x))p(\theta|x)d\theta \tag{4}$$

Note that Bayes risk can be interpreted in a different way:

$$\begin{aligned} R(\delta, \pi) &= \iint L(\theta, \delta(x))p(x|\theta)dx\pi(\theta)d\theta \\ &= \iint L(\theta, \delta(x))p(\theta|x)d\theta \; p(x)dx \\ &= \int r(x, \pi)p(x)d\theta. \end{aligned} \tag{5}$$

The last equation is the posterior risk averaged over the marginal distribution of x. This implies that the Bayes rule can be obtained by taking the Bayes action for each particular x.

Example 1. (Quadratic Loss)

[1] The central idea of Bayesian thinking is to use the data (the posterior).

Define the loss function as $L(\theta, d) = (g(\theta) - d)^2$, known as the *quadratic loss*. The risk function corresponding to the quadratic loss has an interesting decomposition:

$$\begin{aligned} R(\theta, \delta) &= E_\theta (g(\theta) - \delta(X))^2 \\ &= E_\theta \left(g(\theta) - E_\theta \delta(X) - (\delta(X) - E_\theta \delta(X)) \right)^2 \\ &= (g(\theta) - E_\theta \delta(X))^2 + E_\theta (\delta(X) - E_\theta \delta(X))^2 \\ &= [\text{Bias}_\theta(\delta(X))]^2 + \text{Var}_\theta(\delta(X)) \end{aligned} \tag{6}$$

where $\text{Bias}_\theta(\delta(X)) \coloneqq g(\theta) - E_\theta \delta(X)$.[2]

Example 2. (0-1 loss) The 0-1 loss function is defined as follows:

$$L(\theta, d) = \begin{cases} 0 & \text{if } \theta = d \\ k & \text{if } \theta \neq d \end{cases}$$

where $\theta \in \{0, 1\}$, $d \in \{0, 1\}$, and k is some real number. One can think about the 0-1 loss function in the context of hypothesis testing: $\theta = 0$ represents the null hypothesis is true; $\theta = 1$ represents the alternate hypothesis is true; $d = 0$ means that the test procedure fails to reject the null; $d = 1$ means that the test procedure rejects the null.

Example 3. Consider the quadratic loss function: $L(\theta, d) = (\theta - d)^2$. Let $X_1, X_2, ..., X_n$ be independently identically distributed random variable, where $X_i = \theta + \epsilon_i$ with θ being a constant, $E\epsilon_i = 0$ and $E\epsilon_i^2 = \sigma^2$. Let $\delta(X) = \bar{X}$ be the estimator of θ. Note that $\delta(X)$ is unbiased:

$$E_\theta \delta(X) = E_\theta \left(\frac{1}{n} \sum_{i=1}^n X_i \right) = \frac{1}{n} \sum_{i=1}^n E_\theta X_i = \frac{1}{n} \sum_{i=1}^n \theta = \frac{1}{n} n\theta = \theta. \tag{7}$$

The variance of the estimator $\delta(X)$ is:

$$\text{Var}_\theta \delta(X) = E_\theta (\bar{X} - \theta)^2 = \frac{1}{n^2} \sum_{i=1}^n \text{Var}_\theta(X_i) = \frac{\sigma^2}{n}. \tag{8}$$

2 Tower Property

Here is a result that are very useful in this class:

Theorem 4. *(Tower Property)*[3] *Let Y be an integrable random variable and X be a random variable. Then*

$$E\left(E(Y|X)\right) = EY \tag{9}$$

The idea is that "the average of average is an average."

[2] An estimator $\delta(X)$ is said to be *unbiased* if $E_\theta \delta(X) = g(\theta)$.

[3] In Keener it is referred as "smoothing".

3 Exponential Families

Let μ be a measure on $\mathbb{R}^n$, $h : \mathbb{R}^n \to \mathbb{R}$ be a nonnegative function, $T_1, ..., T_s$ are measurable functions mapping from $\mathbb{R}^n$ to $\mathbb{R}$, and $\eta \in \mathbb{R}^s$. Define

$$p_\eta(x) = \exp\left(\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right) h(x) \tag{10}$$

$$A(\eta) = \log \int \exp\left(\sum_{i=1}^{s} \eta_i T_i(x)\right) h(x)\mu(dx) \tag{11}$$

This gives a family of probability densities indexed by η. ($A(\eta)$ ensures that $p_\eta(x)$ is integrated to 1.) The family of densities $\{p_\eta : A(\eta) < \infty\}$ is called an *s-parameter exponential family in canonical form.*

Example 5. (Exponential Distribution) The probability density function of an Exponential(λ) random variable ($\lambda > 0$) is

$$p(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

To show that the density belongs to the exponential family, define μ to be the Lebesgue measure, $h(x) = 1_{(0,\infty)}(x)$ and observe that $\lambda e^{-\lambda x} = \exp(-\lambda x + \log \lambda)$. By setting $\eta = -\lambda$ and $A(\eta) = -\log\left(-\frac{1}{\eta}\right)$, we have the desired form of an exponential family density.

Example 6. (Poisson distribution) The probability density function of an Poisson(λ) random variable ($\lambda > 0$) is

$$p(x) = \begin{cases} \frac{\lambda^x}{x!} e^{-\lambda} & \text{if } x = 0, 1, 2, ... \\ 0 & \text{otherwise} \end{cases}$$

Let μ be the counting measure and rewrite $p(x)$ as

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!} 1_{(0,\infty)}(x) = \frac{1}{x!} \exp(x \log \lambda - \lambda)\, 1_{(0,\infty)}(x) = \exp(\eta x - e^\eta)\left(\frac{1}{x!} 1_{(0,\infty)}(x)\right), \tag{12}$$

where $\eta = \log \lambda$, $A(\eta) = e^\eta$, and $h(x) = \frac{1}{x!} 1_{(0,\infty)}(x)$.

Comment: One may wonder why we are interested in studying exponential families. The main reason is that if the density is in the exponential familiesm we can calculate the expectation (which is an integration) by taking derivatives, which is easy to do in general. Another reason is convenience: exponential families provide a unifying framework for studying certain densities.

More generally, we can set $B(\theta) = A(\eta(\theta))$ and write

$$p_\theta(x) = h(x) \exp\left(\sum_{i=1}^{s} \eta_i(\theta) T_i(x) - B(\theta)\right) \tag{13}$$

Example 7. (Gaussian) The probability density function for a Normal(μ, σ^2) random variable is

$$p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(X-\mu)^2} = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} x^2 - \left(\frac{\mu^2}{2\sigma^2} + \log \sigma^2\right)\right). \tag{14}$$

We have

$$T_1(x) = x; \qquad T_2 = x^2 \tag{15}$$

$$\eta_1(\theta) = \frac{\mu}{\sigma^2}; \quad \eta_2(\theta) = -\frac{1}{2\sigma^2} \tag{16}$$

Example 8. (Random samples) Suppose $X_i \overset{iid}{\sim} p_\theta(x)$ for $i = 1, 2, ..., n$ and $p_\theta(x)$ belongs to the exponential family. The joint probability density function of $X_1, ..., X_n$ is

$$p(x_1, ..., x_n) = \prod_{i=1}^{n} h(x_i) \exp\left(\sum_{i=1}^{s} \eta_i(\theta) \sum_{j=1}^{n} T_i(x_j) - nB(\theta)\right) \tag{17}$$

Theorem 9. *Let* $\Xi_f = \{\eta : \int |f(x)| \exp\left(\sum_i \eta_i T_i(x)\right) h(x)\mu(dx) < \infty\}$. *Then*

$$g(\eta) = \int f(x) \exp\left(\sum_i \eta_i T_i(x)\right) h(x)\mu(dx) \tag{18}$$

is continuous, has continuous partial derivatives of all orders for $\eta \in \Xi_f^o$, *and derivatives can be computed by differentiating under the integral sign.*

Application: Set $f \equiv 1$. Then $g(\eta) = e^{A(\eta)}$ and

$$\begin{aligned}
\frac{\partial}{\partial \eta_j} g(\eta) &= e^{A(\eta)} \frac{\partial A(\eta)}{\partial \eta_j} \\
&= \int \frac{\partial}{\partial \eta_j} \exp\left(\sum_i \eta_i T_i(x)\right) h(x)\mu(dx) \\
&= \int T_j(x) \exp\left(\sum_i \eta_i T_i(x)\right) h(x)\mu(dx)
\end{aligned}$$

which implies that

$$\boxed{\frac{\partial A(\eta)}{\partial \eta_j} = \int T_j(x) p_\eta(x)\mu(dx) = ET_j(x)} \tag{19}$$

4 Moment Generating Functions and Cumulant Generating Functions

The *moment generating function* of a random vector T is defined as

$$M_T(u) = Ee^{u^T T} \tag{20}$$

and the *cumulant generating function* of T is defined as

$$K_T(u) = \log Ee^{u^T T} \tag{21}$$

We could prove that:

$$\alpha_{r_1,...,r_s} \coloneqq E\left(T_1^{r_1} \cdots T_s^{r_s}\right) = \frac{\partial^{r_1}}{\partial u_1^{r_1}} \cdots \frac{\partial^{r_s}}{\partial u_s^{r_s}} M_T(u)\Big|_{u=0} \tag{22}$$

$$\kappa_{r_1,...,r_s} \coloneqq \frac{\partial^{r_1}}{\partial u_1^{r_1}} \cdots \frac{\partial^{r_s}}{\partial u_s^{r_s}} K_T(u)\Big|_{u=0}. \tag{23}$$

Example 10. Let $s = 1$. Then

$$K'_T = \frac{M'_T}{M_T} \tag{24}$$

$$K''_T = \frac{M_T M''_T - M'^2_T}{M_T^2} \tag{25}$$

Setting $u = 0$, we have

$$\kappa_1 = \alpha_1 = ET \tag{26}$$

$$\kappa_2 = ET^2 - (ET)^2 = \mathrm{Var}T \tag{27}$$

4.1 Exponential Family MGF and CGF

For the exponential family, the moment generating function is

$$\begin{aligned} M_T(u) = E_\eta e^{u^T T(x)} &= \int e^{u^T T(x)} e^{\eta^T T(x) - A(\eta)} h(x)\mu(dx) \\ &= e^{A(u+\eta)-A(\eta)} \int e^{(\mu+\eta)^T T(x) - A(u+\eta)} h(x)\mu(dx) \\ &= e^{A(u+\eta)-A(\eta)} \end{aligned} \tag{28}$$

and the cumulant generating function is

$$K_T(u) = A(u+\eta) - A(\eta). \tag{29}$$

Taking derivatives, we have

$$\kappa_{r_1,\ldots,r_s} = \frac{\partial^{r_1}}{\partial \eta_1^{r_1}} \cdots \frac{\partial^{r_s}}{\partial \eta_s^{r_s}} A(\eta) \tag{30}$$

Example 11. (Poisson distribution) Recall that for a Poisson(λ) distribution, $A(\eta) = e^\eta = \lambda$. Hence all the cumulants are $e^\eta = \lambda$ according to (30).