10702/36702 Statistical Machine Learning, Spring 2008: Homework 1 Solutions

February 6, 2008

1 [14 points]

Let Θ be a finite set. Let $L(\theta, \hat{\theta}) = 0$ if $\theta = \hat{\theta}$ and $L(\theta, \hat{\theta}) = 1$ otherwise. Show that the posterior mode is the Bayes estimator.

★ **SOLUTION:**

$$\begin{aligned}
\hat{\theta}_{bayes} &= argmin_{\hat{\theta}} \quad \hat{r}(\theta|x) \\
&= argmin_{\hat{\theta}} \quad \int L(\theta, \hat{\theta})\pi(\theta|x)d\theta \\
&= argmin_{\hat{\theta}} \quad \int I(\theta \neq \hat{\theta})\pi(\theta|x)d\theta \\
&= argmin_{\hat{\theta}} \quad [1 - \int I(\theta = \hat{\theta})\pi(\theta|x)d\theta] \\
&= argmin_{\hat{\theta}} \quad [1 - P(\hat{\theta}|x)] \\
&= argmax_{\hat{\theta}} \quad P(\hat{\theta}|x) \\
&= \text{posterior mean}
\end{aligned}$$

2 [30 points]

Let $X \sim N(\theta, 1)$. Suppose that $\theta \in \Theta = [-C, C]$ where $C = 1/2$. Assume squared error loss.

(a) Verify that $\hat{\theta} = C \tanh(CX)$ is minimax. Hint: Show that $\hat{\theta}$ is the Bayes estimator under the prior $\pi = (1/2)\delta_{-C} + (1/2)\delta_C$ where δ_a denotes a distribution that puts probability 1 at a. You may assume that $R(\theta, \hat{\theta})$ has the following properties: it is continuous, symmetric about 0 and increasing on $[0, c]$.

★ **SOLUTION:** Under the given prior, the posterior probability

$$\pi(\theta = C|X) = \frac{\frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} \exp(-\frac{(X-C)^2}{2})}{\frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} \exp(-\frac{(X-C)^2}{2}) + \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} \exp(-\frac{(X+C)^2}{2})}$$

$$\pi(\theta = -C|X) = \frac{\frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} \exp(-\frac{(X+C)^2}{2})}{\frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} \exp(-\frac{(X-C)^2}{2}) + \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} \exp(-\frac{(X+C)^2}{2})}$$

$$\pi(\theta|X) = 0, \text{ if } \theta \neq C \text{ and } \theta \neq -C$$

With squared error loss, the Bayes estimator is the posterior mean, i.e.

$$\mathbb{E}(\theta|X) = C \cdot \pi(\theta = C|X) - C \cdot \pi(\theta = -C|X) = C\tanh(CX) = \hat{\theta}$$

The Bayes risk $R_\pi(\hat{\theta}) = \pi(\theta = C)R(C,\hat{\theta}) + \pi(\theta = -C)R(-C,\hat{\theta})$. $R(\theta,\hat{\theta}) = \int(\hat{\theta}-\theta)^2\frac{1}{\sqrt{2\pi}}\exp(-\frac{(x-\theta)^2}{2})dx = \int(C\tanh(Cx)-\theta)^2\frac{1}{\sqrt{2\pi}}\exp(-\frac{(x-\theta)^2}{2})dx$. Since $R(\theta,\hat{\theta})$ is continuous, symmetric about 0 and increasing on $[0, C]$, we have $R_\pi(\hat{\theta}) = R(C,\hat{\theta}) > R(\theta,\hat{\theta})$, $\forall\theta \in [-C, C]$. According to Theorem 4, $\hat{\theta} = C\tanh(CX)$ is minimax.

(b) Find the mle (maximum likelihood estimator) $\hat{\theta}$.

★ **SOLUTION:** The mle: $\hat{\theta}_{mle} = X$, if $-C \le X \le C$; $\hat{\theta}_{mle} = C$, if $X > C$; $\hat{\theta}_{mle} = -C$, if $X < -C$.

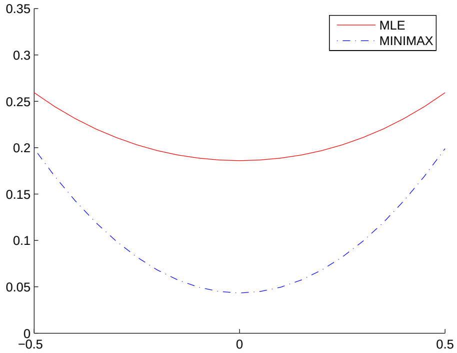
(c) Find the risk of the mle.

★ **SOLUTION:** The risk of the mle:

$$\begin{aligned}
R(\theta,\hat{\theta}_{mle}) = E_\theta((\theta-\hat{\theta}_{mle})^2) &= \int_{-\infty}^{+\infty}(\theta-\hat{\theta}_{mle})^2 f(x;\theta)dx \\
&= \int_{-\infty}^{-C}(\theta+C)^2\frac{1}{\sqrt{2\pi}}\exp(-\frac{(x-\theta)^2}{2})dx + \int_{C}^{+\infty}(\theta-C)^2\frac{1}{\sqrt{2\pi}}\exp(-\frac{(x-\theta)^2}{2})dx \\
&+ \int_{-C}^{C}(\theta-x)^2\frac{1}{\sqrt{2\pi}}\exp(-\frac{(x-\theta)^2}{2})dx \\
&= (\theta+C)^2\Phi(-\theta-C) + (\theta-C)^2\Phi(\theta-C) + \int_{-C}^{C}(\theta-x)^2\frac{1}{\sqrt{2\pi}}\exp(-\frac{(x-\theta)^2}{2})dx
\end{aligned}$$

Where $\Phi(\cdot)$ is the standard normal cdf.

(d) Plot the risk functions of these two estimators.



★ **SOLUTION:**

3 [20 points]

Let $X \sim \text{Binomial}(n, \theta)$.

(a) Find a minimax estimator. Hint: Consider a Bayes estimator based on beta prior.

★ **SOLUTION:** $X \sim \text{Binomial}(n,\theta) \qquad \theta \sim \text{Beta}(\alpha,\beta)$

$$\begin{aligned}
\pi(\theta|x) &\propto f(x|\theta)\pi(\theta) \\
&\propto \binom{n}{x}\theta^x(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&\propto \theta^{\alpha+x-1}(1-\theta)^{n+\beta-x-1} \\
\text{Hence,}\quad \theta|x &\sim \text{Beta}(\alpha+x, n+\beta-x)
\end{aligned}$$

$$\hat{\theta}(x)_{bayes} = E(\pi(\theta|x)) = \frac{\alpha+x}{n+\alpha+\beta} \tag{1}$$

$$\begin{aligned}
R(\theta,\hat{\theta}) &= \text{bias}^2_\theta(\hat{\theta}) + \text{Var}_\theta(\hat{\theta}) \\
&= [E_\theta(\theta - \frac{\alpha+x}{n+\alpha+\beta})]^2 + \text{Var}_\theta(\frac{\alpha+x}{n+\alpha+\beta}) \quad \ldots \quad \text{(from 1)} \\
&= \frac{\theta^2[(\alpha+\beta)^2-n]+\theta[n-2\alpha(\alpha+\beta)]+\alpha^2}{(n+\alpha+\beta)^2}
\end{aligned}$$

$R(\theta,\hat{\theta})$ is constant in θ if $(\alpha+\beta)^2 = n$ and $2\alpha(\alpha+\beta) = n$.
Solving these equations we get the following values:

$$\alpha = \beta = \frac{\sqrt{n}}{2} \tag{2}$$

Substituting eq. 2 into eq. 1, we get

$$\hat{\theta}(x)_{minimax} = \frac{x+\frac{\sqrt{n}}{2}}{n+\sqrt{n}}$$

(b) Plot the risk of the minimax estimator, the mle and the Bayes estimator using a flat prior, for $n = 5, 50, 100$.

★ **SOLUTION:** Minimax Risk: $R(\theta,\hat{\theta}_{minimax}) = \frac{\alpha^2}{(n+\alpha+\beta)^2} = \frac{1}{4(\sqrt{n}+1)^2}$
MLE estimate:

$$\hat{\theta}_{mle} = \frac{x}{n}$$

MLE Risk:

$$\begin{aligned}
R(\theta,\hat{\theta}_{mle}) &= \text{bias}^2_\theta(\hat{\theta}) + \text{Var}_\theta(\hat{\theta}) \\
&= (\theta - \frac{E(x)}{n})^2 + \text{Var}_\theta(\frac{x}{n}) \\
&= (\theta - \frac{n\theta}{n})^2 + \frac{1}{n^2}n\theta(1-\theta) \\
&= \frac{\theta(1-\theta)}{n}
\end{aligned}$$

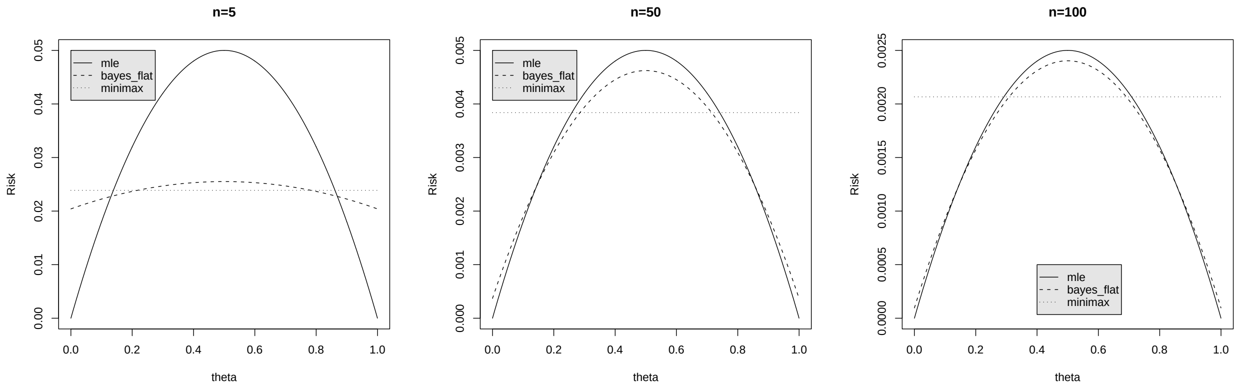
Bayes estimate with flat prior is equivalent to $Beta(1,1)$ prior. Hence using eq. 1,

$$\hat{\theta}_{bayes} = \frac{x+1}{n+2}$$

.

Bayes Risk:

$$\begin{aligned}
R(\theta, \hat{\theta}_{bayes}) &= \mathsf{bias}^2_\theta(\hat{\theta}) + \mathsf{Var}_\theta(\hat{\theta}) \\
&= (E_\theta(\theta - \frac{x+1}{n+2}))^2 + \mathsf{Var}_\theta(\frac{x+1}{n+2}) \\
&= \frac{n\theta(1-\theta) + (2\theta-1)^2}{(n+2)^2}
\end{aligned}$$



4 [36 points]

This question will help you explore the differences between Bayesian and frequentist inference. Let $X_1, \ldots, X_n$ be a sample from a multivariate normal distribution with mean $\mu = (\mu_1, \ldots, \mu_p)^T$ and covariance matrix equal to the identity matrix I. Note that each X_i is a vector of length p.

The following facts will be helpful. If $Z_1, \ldots, Z_k$ are independent $N(0,1)$ and $a_1, \ldots, a_k$ are constants, then we say that $Y = \sum_{j=1}^k (Z_j + a_j)^2$ has a non-central χ^2 distribution with k degrees of freedom and noncentrality parameter $\|a\|^2$. The mean and variance of Y are $k + \|a\|^2$ and $2k + 4\|a\|^2$.

(a) Find the posterior under the improper prior $\pi(\mu) = 1$.

★ SOLUTION:

$$\begin{aligned}
\pi(\mu|X_1, \ldots, X_n) &\propto L_n(\mu)\pi(\mu) \\
&= \prod_{i=1}^n [(2\pi)^{-p/2} \exp(-\frac{1}{2}(x_i - \theta)^T (x_i - \theta))] \\
&\propto \exp(-\frac{n}{2}(\mu - \bar{x})^T(\mu - \bar{x}))
\end{aligned}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$. Therefore, the posterior of μ is multivariate Gaussian with mean $\bar{x}$ and covariance matrix $\frac{1}{n} I_{p \times p}$.

(b) Let $\theta = \sum_{j=1}^p \mu_j^2$. Our goal is to learn θ. Find the posterior for θ. Express your answers in terms of noncentral χ^2 distributions. Find the posterior mean $\tilde{\theta}$.

★ **SOLUTION:** $\forall j \in [1,\ldots,p]$, $\pi(\sqrt{n}\mu_j|X_1,\ldots,X_n) \sim N(\sqrt{n}\bar{x}_j, 1)$, where $\bar{x}_j = \frac{1}{n}\sum_{i=1}^n x_{ij}$. Therefore, $\pi(n\theta|X_1,\ldots,X_n) \sim \chi^2_p(n\|\bar{x}\|^2)$, and the posterior mean $\tilde{\theta} = \frac{1}{n}\mathbb{E}(n\theta|X_1,\ldots,X_n) = \frac{1}{n}(p+n\|\bar{x}\|^2) = \frac{p}{n} + \|\bar{x}\|^2$.

(c) The usual frequentist estimator is $\hat{\theta} = \|\bar{X}\|^2 - p/n$. Show that, for any n,

$$\frac{\mathbb{E}_\theta\|\theta - \tilde{\theta}\|^2}{\mathbb{E}_\theta\|\theta - \hat{\theta}\|^2} \to \infty$$

as $p \to \infty$.

★ **SOLUTION:** $\forall j \in [1,\ldots,p]$, $\bar{X}_j = \frac{1}{n}\sum_{i=1}^n X_{ij}$. It is easy to see that $\bar{X}_j \sim N(\mu_j, 1/n)$, and $\sqrt{n}\bar{X}_j \sim N(\sqrt{n}\mu_j, 1)$. Therefore, $n\|\bar{X}\|^2 \sim \chi^2_p(n\|\mu\|^2)$, $\mathbb{E}_\theta\|\bar{X}\|^2 = \frac{p}{n} + \|\mu\|^2 = \frac{p}{n} + \theta$, and $\mathbb{V}_\theta(\|\bar{X}\|^2) = \frac{2p}{n^2} + \frac{4\|\mu\|^2}{n} = \frac{2p}{n^2} + \frac{4\theta}{n}$.

$\mathbb{E}_\theta\|\theta - \tilde{\theta}\|^2 = (\theta - \mathbb{E}_\theta\tilde{\theta})^2 + \mathbb{V}_\theta(\tilde{\theta}) = (\mathbb{E}_\theta\|\bar{X}\|^2 + \frac{p}{n} - \theta)^2 + \mathbb{V}_\theta(\|\bar{X}\|^2) = \frac{4p^2}{n^2} + \frac{2p}{n^2} + \frac{4\theta}{n}$.

$\mathbb{E}_\theta\|\theta - \hat{\theta}\|^2 = (\theta - \mathbb{E}_\theta\hat{\theta})^2 + \mathbb{V}_\theta(\hat{\theta}) = (\mathbb{E}_\theta\|\bar{X}\|^2 - \frac{p}{n} - \theta)^2 + \mathbb{V}_\theta(\|\bar{X}\|^2) = \frac{2p}{n^2} + \frac{4\theta}{n}$.

Therefore, for any n, as $p \to \infty$

$\frac{\mathbb{E}_\theta\|\theta-\tilde{\theta}\|^2}{\mathbb{E}_\theta\|\theta-\hat{\theta}\|^2} = \frac{\frac{4p^2}{n^2}+\frac{2p}{n^2}+\frac{4\theta}{n}}{\frac{2p}{n^2}+\frac{4\theta}{n}} \to \infty$

(d) Repeat the analysis with a $N(0, \tau^2 I)$ prior.

★ **SOLUTION:** By similar analysis, we have $\pi(\mu|X_1,\ldots,X_n) \sim N(\frac{n}{n+\frac{1}{\tau^2}}\bar{x}, \frac{1}{n+\frac{1}{\tau^2}}I_{p\times p})$, $\pi((n + \frac{1}{\tau^2})\theta|X_1,\ldots,X_n) \sim \chi^2_p(\frac{n^2}{n+\frac{1}{\tau^2}}\|\bar{x}\|^2)$. Therefore, the posterior mean $\tilde{\theta} = \frac{p}{n+\frac{1}{\tau^2}} + \frac{n^2}{(n+\frac{1}{\tau^2})^2}\|\bar{x}\|^2$.

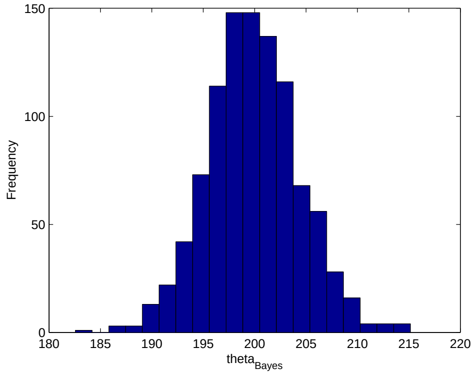
$\mathbb{E}_\theta\|\theta - \tilde{\theta}\|^2 = (\theta - \mathbb{E}_\theta\tilde{\theta})^2 + \mathbb{V}_\theta(\tilde{\theta}) = (\mathbb{E}_\theta(\frac{n^2}{(n+\frac{1}{\tau^2})^2}\|\bar{X}\|^2) + \frac{p}{n+\frac{1}{\tau^2}} - \theta)^2 + \mathbb{V}_\theta(\frac{n^2}{(n+\frac{1}{\tau^2})^2}\|\bar{X}\|^2) = (\frac{2n+\frac{1}{\tau^2}}{(n+\frac{1}{\tau})^2}p - \frac{\frac{2n}{\tau^2}+\frac{1}{\tau^4}}{(n+\frac{1}{\tau})^2}\theta)^2 + \frac{n^2}{(n+\frac{1}{\tau^2})^4}(2p + 4n\theta)$.

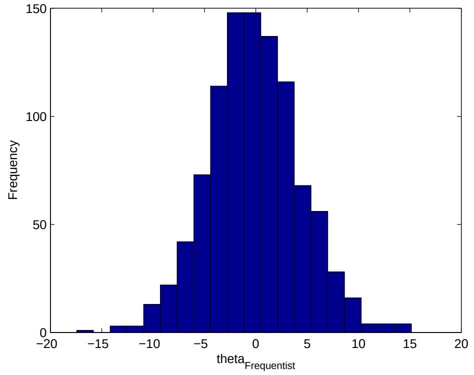
Therefore, for any n, as $p \to \infty$

$\frac{\mathbb{E}_\theta\|\theta-\tilde{\theta}\|^2}{\mathbb{E}_\theta\|\theta-\hat{\theta}\|^2} = \frac{(\frac{2n+\frac{1}{\tau^2}}{(n+\frac{1}{\tau})^2}p - \frac{\frac{2n}{\tau^2}+\frac{1}{\tau^4}}{(n+\frac{1}{\tau})^2}\theta)^2 + \frac{n^2}{(n+\frac{1}{\tau^2})^4}(2p+4n\theta)}{\frac{2p}{n^2}+\frac{4\theta}{n}} \to \infty$

(e) Set $n = 10$, $p = 1000$, $\theta = (0,\ldots,0)^T$. Simulate (in R) data N times, with $N = 1000$. Draw a histogram of the Bayes estimator (with flat prior) and the frequentist estimator.

★ **SOLUTION:** The histograms are as follows.





(f) Interpret your findings.

★ **SOLUTION:** From the figures, we can see that the two histograms have the save shape, and the frequentist estimator has less bias compared with the Bayes estimator. According to (c), we can see that the frequentist estimator ($\hat{\theta} = \|\bar{X}\|^2 - \frac{p}{n}$) is unbiased $\theta - \mathbb{E}_\theta \hat{\theta} = \theta$, whereas the Bayes estimator with flat prior ($\tilde{\theta} = \|\bar{X}\|^2 + \frac{p}{n}$) is biased, $\theta - \mathbb{E}_\theta \tilde{\theta} = \frac{2p}{n}$. This bias is significant when p is much larger than n.