

10-702 Statistical Machine Learning: Assignment 1 Solutions

1. Review of Maximum Likelihood.

Let X_1, \dots, X_n be a random sample where $X_i \in \{1, 2, \dots, k\}$. Let $\theta \in [0, 1]$ and suppose that $\mathbb{P}(X_i = 1) = \theta$ and $\mathbb{P}(X_i = j) = \bar{\theta}$ for $j > 1$ where $\bar{\theta} = (1 - \theta)/(k - 1)$.

- (a) Find the mle $\hat{\theta}$. Let $Y = 1$ when $X = 1$ and $Y = 0$ otherwise. $P(Y = 1) = \theta$ and $P(Y = 0) = 1 - \theta$. Hence, $P \sim \text{Bernoulli}(\theta)$

$$L(\theta) = \prod_{i=1}^n \theta^{Y_i} (1 - \theta)^{1 - Y_i}$$

$$\text{Solving } \frac{\partial}{\partial \theta} l(\theta) = 0 \text{ we get, } \hat{\theta} = \frac{1}{n} \sum_{i=1}^n Y_i = \frac{1}{n} \sum_{i=1}^n I(X_i = 1)$$

- (b) Find the Fisher information.

$$I_n(\theta) = -\mathbb{E}\left[\frac{\partial^2 l(\theta)}{\partial \theta^2}\right]$$

Again assuming the transformation to Y ,

$$\begin{aligned} I_n(\theta) &= -\mathbb{E}\left[\frac{\partial}{\partial \theta^2} (\log(\theta) \sum Y_i + \log(1 - \theta) \sum (1 - Y_i))\right] \\ &= -\mathbb{E}\left[-\frac{1}{\theta^2} \sum Y_i - \frac{1}{(1 - \theta)^2} \sum (1 - Y_i)^2\right] \\ &= \frac{n}{\theta} + \frac{n}{1 - \theta} \quad \because E(Y_i) = \theta \\ &= \frac{n}{\theta(1 - \theta)} \end{aligned}$$

- (c) Find an approximate $1 - \alpha$ confidence interval for θ .

$$C_n = (\hat{\theta} - z_{\alpha/2} se, \hat{\theta} + z_{\alpha/2} se) \text{ where } se = \sqrt{\frac{1}{I_n(\hat{\theta})}}$$

Here substitute $I_n(\theta)$ using (2) and $\hat{\theta}$ from (1). $Z \sim N(0, 1)$.

- (d) Find the bias and variance of $\hat{\theta}$.

$$\begin{aligned} \text{bias}(\hat{\theta}) &= \mathbb{E}(\hat{\theta}) - \theta \\ &= \mathbb{E}\left(\frac{1}{n} \sum Y_i\right) - \theta \\ &= \theta - \theta = 0 \end{aligned}$$

Now, $\hat{\theta} = \frac{1}{n} \sum Y_i$ where $Y_i \sim \text{Bernoulli}(\theta)$. Hence, $\sum_{i=1}^n Y_i \sim \text{Binominal}(n, \theta)$.

$$\begin{aligned} \text{Var}(\hat{\theta}) &= \frac{1}{n^2} \text{Var}\left(\sum Y_i\right) \\ &= \frac{1}{n^2} n\theta(1 - \theta) \\ &= \frac{\theta(1 - \theta)}{n} \end{aligned}$$

(e) Show that $\hat{\theta}$ is consistent.

$$\text{MSE} = \text{bias}^2(\hat{\theta}) + \text{Var}(\hat{\theta}) = \frac{\theta(1-\theta)}{n}$$

$\lim_{n \rightarrow \infty} \text{MSE} = 0$. Hence, $\hat{\theta}$ is consistent.

2. Probability.

Let X_1, \dots, X_n be iid and assume that $-1 \leq X_i \leq 1$. Also assume that X_i has mean 0.

(a) Use Hoeffding's inequality to show that $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ is close to 0 with high probability.

For X with mean 0, and bounded between $[-1, 1]$, Hoeffding's inequality says that

$$P(|\bar{X}_n| > \epsilon) \leq 2 \exp\left(\frac{-2n^2\epsilon^2}{\sum_{i=1}^n 4}\right)$$

Thus,

$$P(|\bar{X}_n| > \epsilon) \leq 2 \exp(-n\epsilon^2/2)$$

If you assume that ϵ is fixed, then clearly, this probability is going down to 0 exponentially fast in n , and thus as $n \rightarrow \infty$, \bar{X}_n is close to 0 with high probability.

However, usually, as n increases, we expect that our estimate \bar{X}_n will get more accurate, and thus, we expect that as n increases, ϵ reduces, i.e. accuracy increases.

Let δ be the probability of error, i.e

$$P(|\bar{X}_n| > \epsilon) \leq 2 \exp(-n\epsilon^2/2) = \delta$$

Then, with high probability $1 - \delta$, \bar{X}_n is close to zero, when ϵ is chosen as a function of n and δ as

$$2 \exp(-n\epsilon^2/2) = \delta$$

$$\epsilon = \left(\frac{2}{n} \log\left(\frac{2}{\delta}\right)\right)^{\frac{1}{2}}$$

This allows us to claim that \bar{X}_n is close to zero with high probability $1 - \delta$ even when ϵ reduces as a function of n .

(b) Show that there exists c such that

$$\frac{1}{n} \sum_{i=1}^n X_i^2 \xrightarrow{P} c$$

and find c .

Let $Y = X^2$. Then

$$\frac{1}{n} \sum_{i=1}^n X_i^2 = \frac{1}{n} \sum_{i=1}^n Y_i \xrightarrow{P} E(Y) = E(X^2) = \text{Var}(X) + [E(X)]^2 = \text{Var}(X)$$

(c) Say whether the following statements are true or false and explain why.

i. $\bar{X}_n = o(1)$.

FALSE. \bar{X}_n converges in probability to 0, but this is not convergence everywhere. Specifically, we can construct a series of X_1, \dots, X_n whose sample mean does not go to zero as $n \rightarrow \infty$.

ii. $\bar{X}_n = o_P(1)$.

TRUE. \bar{X}_n converges in probability to 0, by Hoeffding's inequality.

iii. $\bar{X}_n = o_P(n)$.

TRUE. $\bar{X}_n = o_P(1)$ implies $\bar{X}_n = o_P(n)$

iv. $\bar{X}_n = o_P(1/n)$.

FALSE. $n\bar{X}_n = \sum_i X_i$ which is not bounded in probability.

v. $\bar{X}_n = O_P(n^{-1/2})$.

TRUE. Central Limit Theorem

vi. $\bar{X}_n = O_P(n^{-1})$.

FALSE. No known result that says this.

3. This question will help you explore the differences between Bayesian and frequentist inference. Let X_1, \dots, X_n be a sample from a multivariate Normal distribution with mean $\mu = (\mu_1, \dots, \mu_p)^T$ 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, \dots, Z_k are independent $N(0, 1)$ and a_1, \dots, 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 $\pi(\mu) = 1$.

$$\begin{aligned}
\pi(\mu|X) &\propto \prod_{i=1}^n P(X_i|\mu)\pi(\mu) \\
&\propto \prod_{i=1}^n \exp\left(-\frac{1}{2}(X_i - \mu)^T(X_i - \mu)\right) \\
&= \prod_{i=1}^n \exp\left(-\frac{1}{2}(X_i^T X_i - 2X_i^T \mu + \mu^T \mu)\right) \\
&= \exp\left(-\frac{1}{2}\left(\sum_{i=1}^n X_i^T X_i - 2\sum_{i=1}^n X_i^T \mu + n\mu^T \mu\right)\right) \\
&= \exp\left(-\frac{1}{2}n\left(\frac{1}{n}\sum_{i=1}^n X_i^T X_i - 2\bar{X}\mu + \mu^T \mu\right)\right)
\end{aligned}$$

completing the square around μ

$$\begin{aligned}
&= \exp\left(-\frac{1}{2}n\left((\mu - \bar{X})^T(\mu - \bar{X}) + \frac{1}{n}\sum_{i=1}^n X_i^T X_i - \bar{X}^T \bar{X}\right)\right) \\
&\propto \exp\left(-\frac{1}{2}((\mu - \bar{X})^T n(\mu - \bar{X}))\right) \\
&\sim N(\bar{X}, \frac{1}{n}I)
\end{aligned}$$

- (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 $\tilde{\theta}$.

We have

$$\theta = \sum_{i=1}^p \mu_i^2$$

We also have $\mu_i|X \sim N(\bar{X}_i, \frac{1}{n})$ implies $(\sqrt{n}\mu_i)|X \sim N(\sqrt{n}\bar{X}_i, 1) = Z_i + \sqrt{n}\bar{X}_i$

$$\begin{aligned}
\theta &= \frac{1}{n} \sum_{i=1}^p (\sqrt{n}\mu_i)^2 \\
&= \frac{1}{n} \sum_{i=1}^p (Z_i + \sqrt{n}\bar{X}_i)^2 \\
n\theta &= \sum_{i=1}^p (Z_i + \sqrt{n}\bar{X}_i)^2
\end{aligned}$$

Then $n\theta$ is distributed as a noncentral χ^2 distribution with p degrees of freedom and non-central parameter $n \|\bar{X}\|^2$

the mean of $n\theta$ is $p + n \|\bar{X}\|^2$. Therefore $\tilde{\theta} = \frac{p+n\|\bar{X}\|^2}{n} = \frac{p}{n} + \|\bar{X}\|^2$

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

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

as $p \rightarrow \infty$.

Lets consider the bias of $\hat{\theta}$.

$$\begin{aligned} E(\hat{\theta}) - \theta &= E(|\bar{X}_n|^2) - \frac{p}{n} - \theta \\ &= \sum E(\bar{X}_i^2) - \frac{p}{n} - \theta \\ &= \sum (Var(\bar{X}_i) - E(\bar{X}_i)^2) - \frac{p}{n} - \theta \\ &= \sum \left(\frac{1}{n} - u_i^2 \right) - \frac{p}{n} - \theta \\ &= \frac{p}{n} + \sum u_i^2 - \frac{p}{n} - \theta \\ &= 0 \end{aligned}$$

From this, we can also tell that the bias of $\tilde{\theta} = 2\frac{p}{n}$

Lets consider the variance of $\hat{\theta}$. First we note that the $Variance(\hat{\theta}) = Variance(\tilde{\theta})$

$$\begin{aligned} Variance(\tilde{\theta}) &= Variance(|\bar{X}_n|^2) \\ &= Variance\left(\sum \bar{X}_i^2\right) \\ &\quad (\bar{X}_i's \text{ are independent}) \\ &= \sum_{i=1}^p Variance(\bar{X}_i^2) \\ &\quad (\text{since } X \text{ has a covariance matrix of } I) \\ &\quad (\text{all terms have the same variance}) \\ &= p \times (Variance(\bar{X}_1^2)) \end{aligned}$$

We note that $Variance(\bar{X}_1^2)$ is independent of p (whatever it really is). and we let this value be k , therefore writing $Variance(\tilde{\theta}) = Variance(\hat{\theta}) = pk$.

Therefore $\hat{\theta}$ is an unbiased estimator

$$\begin{aligned}
\lim_{p \rightarrow \infty} \frac{E|\theta - \tilde{\theta}|^2}{E|\theta - \hat{\theta}|^2} &= \lim_{p \rightarrow \infty} \frac{\text{Variance}(\tilde{\theta}) + \text{Bias}(\tilde{\theta})^2}{\text{Variance}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2} \\
&= \lim_{p \rightarrow \infty} \frac{pk + 4\frac{p^2}{n^2}}{pk} \\
&= \lim_{p \rightarrow \infty} \frac{k + 4\frac{p}{n^2}}{k} \\
&= \lim_{p \rightarrow \infty} 1 + 4\frac{p}{n^2k} \\
&= \infty
\end{aligned}$$

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

$$\pi(\mu) \sim N(0, \tau^2 I)$$

$$\begin{aligned}
\pi(\mu|X) &\propto \prod_{i=1}^n P(X_i|\mu)\pi(\mu) \\
&\propto \exp\left(-\frac{1}{2} \frac{1}{\tau^2} \mu^T \mu\right) \prod_{i=1}^n \exp\left(-\frac{1}{2} (X_i - \mu)^T (X_i - \mu)\right) \\
&= \exp\left(-\frac{1}{2} \frac{1}{\tau^2} \mu^T \mu\right) \prod_{i=1}^n \exp\left(-\frac{1}{2} (X_i^T X_i - 2X_i^T \mu + \mu^T \mu)\right) \\
&= \exp\left(-\frac{1}{2} \left(\sum_{i=1}^n X_i^T X_i - 2\sum_{i=1}^n X_i^T \mu + (n + \frac{1}{\tau^2})\mu^T \mu\right)\right) \\
&\propto \exp\left(-\frac{1}{2} \left(n + \frac{1}{\tau^2}\right) \left(\mu^T \mu - 2\left(\frac{n}{n + \tau^{-2}}\right) \bar{X} \mu\right)\right) \\
&\text{completing the square around } \mu \\
&= \exp\left(-\frac{1}{2} \left(n + \frac{1}{\tau^2}\right) \left(\mu - \frac{n\bar{X}}{n + \tau^{-2}}\right)^T \left(\mu - \frac{n\bar{X}}{n + \tau^{-2}}\right)\right) \\
&\sim N\left(\frac{n\bar{X}}{n + \tau^{-2}}, \frac{\tau^2}{n\tau^2 + 1} I\right)
\end{aligned}$$

Using the same argument, $\tilde{\theta}' = \frac{\tau^2}{n\tau^2 + 1} \left(p + \frac{n^2 \tau^2}{n\tau^2 + 1} \|\bar{X}\|\right)$

Continuing from **3c**,

Now, we have $\text{Variance}(\tilde{\theta}') = \left(\frac{n\tau^2}{n\tau^2 + 1}\right)^4 pk$.

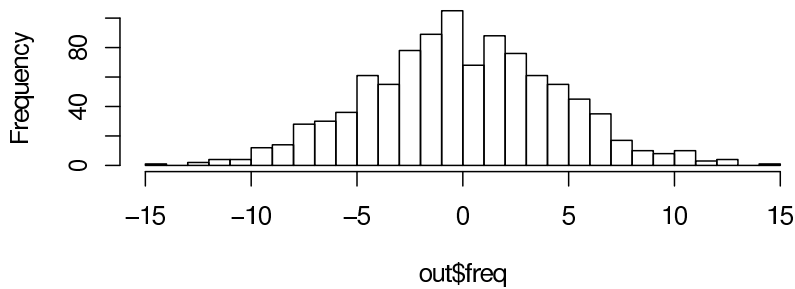
And

$$\begin{aligned} \text{Bias}(\tilde{\theta}') &= E(\tilde{\theta}') - \theta \\ &= \frac{\tau^2}{n\tau^2 + 1}p + \frac{n\tau^4}{(n\tau^2 + 1)^2}p + \left[\left(\frac{n\tau^2}{n\tau^2 + 1} \right)^2 - 1 \right] \theta \end{aligned}$$

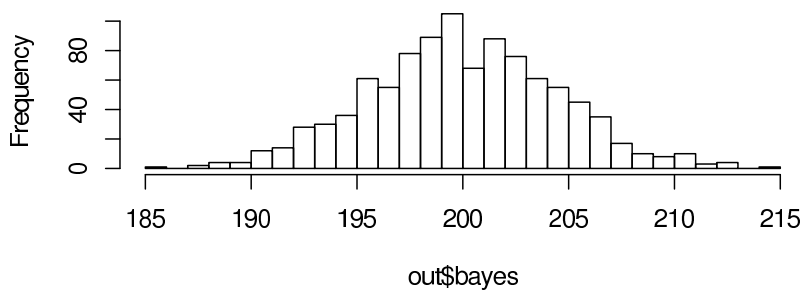
$$\begin{aligned} \lim_{p \rightarrow \infty} \frac{E|\theta - \tilde{\theta}'|^2}{E|\theta - \hat{\theta}|^2} &= \lim_{p \rightarrow \infty} \frac{\text{Variance}(\tilde{\theta}') + \text{Bias}(\tilde{\theta}')^2}{\text{Variance}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2} \\ &= \infty \end{aligned}$$

- (e) Set $n = 10$, $p = 1000$, $\mu = (0, \dots, 0)^T$. Simulate (in R) data N times, with $N = 1000$. Draw a histogram of the Bayes estimator (with flat prior) and the frequentist estimator. (R code for this question can be found on the web site.)

Histogram of out\$freq



Histogram of out\$bayes



- (f) Interpret your findings.

The Bayes estimator and the frequentist estimator are extremely far apart. The true value of θ should be $\theta = 0$, but the Bayes estimator was unable to obtain that. As we calculated, the Bayes estimator has a bias of $\frac{2p}{n}$ which matches the plot.

4. Minimality and Bayes.

Let $X_1, \dots, X_n \sim \text{Bernoulli}(p)$. In what follows we use squared error loss.

- (a) Find the mle \hat{p} . Find the bias, variance and risk (mean squared error) $R(p, \hat{p})$ of \hat{p} .
From question 1, we know that $\hat{p} = \bar{X}_n$.

$$\text{Bias} = p - E(\bar{X}_n) = 0$$

$$\text{Variance} = \frac{1}{n^2} \sum_i \text{Var}(X_i) = \frac{p(1-p)}{n}$$

$$\text{Risk} = \text{Variance} + \text{Bias}^2 = \frac{p(1-p)}{n}$$

- (b) Recall that a Beta(α, β) density has the form

$$\pi(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1} \propto p^{\alpha-1} (1-p)^{\beta-1}.$$

Let p have a Beta(v, v) prior. Find the Bayes estimator \bar{p} . Find the bias, variance and risk $R(p, \bar{p})$ of \bar{p} .

Since the Beta distribution is conjugate prior for Bernoulli, the posterior

$$p|X \sim \text{Beta}(n\bar{X}_n + v, n - n\bar{X}_n + v)$$

The Bayes Estimator \bar{p} is the mean of the posterior distribution. Thus,

$$\bar{p} = \frac{n\bar{X}_n + v}{n + 2v}$$

$$\text{Bias} = E\left(\frac{n\bar{X}_n + v}{n + 2v}\right) - p = \frac{np + v}{n + 2v} - p = \frac{v(1 - 2p)}{n + 2v}$$

$$\text{Variance} = \text{Var}\left(\frac{n\bar{X}_n + v}{n + 2v}\right) = \frac{n^2}{(n + 2v)^2} \text{Var}(\bar{X}_n) + 0 = \frac{np(1-p)}{(n + 2v)^2}$$

$$\text{Risk} = \text{Variance} + \text{Bias}^2 = \frac{v^2(1 - 2p)^2 + np(1-p)}{(n + 2v)^2}$$

- (c) Show that $R(p, \bar{p})$ is constant (as a function of p) if v is chosen appropriately. Since \bar{p} is a Bayes estimator and has constant risk, it is the minimax estimator.

$$R(p, \bar{p}) = \frac{v^2(1 - 2p)^2 + np(1-p)}{(n + 2v)^2}$$

Substitute $v = \sqrt{(n/4)}$ to get

$$R(p, \bar{p}) = \frac{n/4(1 - 4p + 4p^2) + np(1-p)}{(n + 2\sqrt{n}/2)^2}$$

$$R(p, \bar{p}) = \frac{n/4 - np + np^2 + np - np^2}{(n + \sqrt{n})^2}$$

$$R(p, \bar{p}) = \frac{n}{4(n + \sqrt{n})^2}$$

which is a constant (as a function of p), thus, \bar{p} is the minimax estimator.