ST451 Bayesian Machine Learning

Exercises

1. Let $y = (y_1, \ldots, y_n)$ be a random sample from a $\mathrm{N}(\theta, \sigma^2)$ distribution with σ^2 known.

   (a) Show that the likelihood is proportional to

   $$f(y|\theta) \propto \exp\left(-\frac{n(\bar{y}-\theta)^2 + (n-1)S^2}{2\sigma^2}\right).$$

   where $\bar{x}$ is the sample mean and S^2 is the sample variance

   $$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

   Hence the likelihood simplifies to

   $$f(y|\theta) \propto \exp\left(-\frac{(\theta-\bar{y})^2}{2\frac{\sigma^2}{n}}\right)$$

   Answer: The joint density of the sample y is

   $$\begin{aligned} f(y|\theta,\sigma^2) = f(y_1,\ldots,y_n|\theta,\sigma^2) &= \prod_{i=1}^{n}(2\pi\sigma^2)^{-1/2}\exp\left(-\frac{(y_i-\theta)^2}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{\sum_{i=1}^{n}(y_i-\theta)^2}{2\sigma^2}\right) \end{aligned}$$

   Hence, it suffices to show that

   $$\sum_{i=1}^{n}(y_i-\theta)^2 = n(\bar{y}-\theta)^2 + \sum_{i=1}^{n}(y_i-\bar{y})^2.$$

   Note that

   $$\sum_{i=1}^{n}(y_i-\theta)^2 = \sum_{i=1}^{n}(y_i^2 - 2\theta y_i + \theta^2) = \sum_{i=1}^{n} y_i^2 - 2\theta n\bar{y} + n\theta^2$$

   $$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i^2 - 2\bar{y}y_i + \bar{y}^2) = \sum_{i=1}^{n} y_i^2 - 2n\bar{y}^2 + n\bar{y}^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

   Subtracting the second of these equations from the first yields

   $$\sum_{i=1}^{n}(y_i-\theta)^2 - \sum_{i=1}^{n}(y_i-\bar{y})^2 = n\bar{y}^2 - 2\theta n\bar{y} + n\theta^2 = n(\bar{y}-\theta)^2$$

   Since σ^2 is known, we are interested in $f(y|\theta)$ which is proportional to

$$f(y|\theta) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{n(\bar{y}-\theta)^2 + (n-1)S^2}{2\sigma^2}\right)$$
$$\propto \exp\left(-\frac{n(\bar{y}-\theta)^2}{2\sigma^2}\right) \exp\left(-\frac{(n-1)S^2}{2\sigma^2}\right)$$
$$\propto \exp\left(-\frac{(\theta-\bar{y})^2}{2\frac{\sigma^2}{n}}\right)$$

(b) Set the prior for θ to be $\mathrm{N}(\mu, \tau^2\sigma^2)$ and derive its posterior distribution. (You can use the above result)
Answer: The prior for θ is set to be $\mathrm{N}(\mu, \tau^2\sigma^2)$. Hence we can write

$$\pi(\theta) \propto \exp\left(-\frac{(\theta-\mu)^2}{2\tau^2\sigma^2}\right).$$

Using the result of part (i), the posterior is then proportional to

$$\pi(\theta|y) \propto f(y|\theta)\pi(\theta) \propto \exp\left(-\frac{(\theta-\bar{y})^2}{2\frac{\sigma^2}{n}}\right) \exp\left(-\frac{(\theta-\mu)^2}{2\sigma^2\tau^2}\right)$$
$$= \exp\left(-\frac{\theta^2+2\theta\bar{y}}{2\frac{\sigma^2}{n}} - \frac{\theta^2-2\theta\mu}{2\sigma^2\tau^2}\right) \exp\left(-\frac{\bar{y}^2}{2\frac{\sigma^2}{n}}\right) \exp\left(-\frac{\mu^2}{2\sigma^2\tau^2}\right)$$
$$\propto \exp\left(-\frac{\theta^2+2\theta\bar{y}}{2\frac{\sigma^2}{n}} - \frac{\theta^2-2\theta\mu}{2\sigma^2\tau^2}\right)$$
$$= \exp\left(-\frac{\tau^2\theta^2 - 2\theta\bar{y}\tau^2 + \frac{1}{n}\theta^2 - 2\theta\mu\frac{1}{n}}{2\frac{\sigma^2}{n}\tau^2}\right)$$
$$= \exp\left(-\frac{(\frac{1}{n}+\tau^2)\theta^2 - 2\theta(\bar{y}\tau^2 + \mu\frac{1}{n})}{2\frac{\sigma^2}{n}\tau^2}\right)$$
$$= \exp\left(-\frac{\theta^2 - 2\theta\frac{\bar{y}\tau^2+\mu\frac{1}{n}}{(\frac{1}{n}+\tau^2)}}{2\frac{\frac{\sigma^2}{n}\tau^2}{(\frac{1}{n}+\tau^2)}}\right) \overset{\mathcal{D}}{=} \mathrm{N}\left(\frac{\frac{1}{n}\mu + \tau^2\bar{y}}{\tau^2+\frac{1}{n}}, \frac{\tau^2\frac{\sigma^2}{n}}{\tau^2+\frac{1}{n}}\right) \overset{\mathcal{D}}{=} \mathrm{N}\left(\frac{\frac{1}{\tau^2}\mu + n\bar{y}}{\frac{1}{\tau^2}+n}, \frac{\sigma^2}{\frac{1}{\tau^2}+n}\right)$$

2. Suppose that $y_i \sim N(\mu, 1)$ for $i = 1, \ldots, n$ and that the y_i's are independent.

(a) Show that the sample mean estimator $\hat{\mu}_1 = \frac{1}{n}\sum_i^n y_i$ is obtained from minimising the least squares criterion

$$\hat{\mu}_1 = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \mu)^2,$$

and that $\hat{\mu}_1$ an unbiased estimator of μ. Also find the variance of $\hat{\mu}_1$.
Answer: Show that the derivative of $\sum_{i=1}^n (y_i-\mu)^2$ wrt μ is equal to $2\sum_i y_i - 2n\mu$. Setting it equal to 0 and solving then yields $\hat{\mu}_1 = \frac{1}{n}\sum_i^n y_i$. We then get

$$E(\frac{1}{n}\sum_{i}^{n} y_i) = \frac{1}{n}\sum_{i=1}^{n} E(y_i) = \mu,$$

which implies that the estimator is unbiased. For the variance not that

$$\mathrm{var}(\frac{1}{n}\sum_{i}^{n} y_i) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}(y_i) = \frac{1}{n}$$

(b) Consider adding a penalty term to the least squares criterion, and therefore using the estimator that minimises

$$\hat{\mu}_2 = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \mu)^2 + \lambda \mu^2$$

for the mean, where λ is a non-negative tuning parameter. Derive $\hat{\mu}_2$, find it bias and show that its variance is lower than that of $\hat{\mu}_1$

Answer: The derivative w.r.t. μ is $2\sum(y_i - \mu) + 2\lambda\mu$. Setting it to 0 gives

$$\hat{\mu}_2 = \frac{\sum_{i=1}^{n} y_i}{n + \lambda}.$$

Then

$$E(\hat{\mu}_2) = \frac{n}{n+\lambda}\mu,$$

$$\text{Bias}(\hat{\mu}_2) = E(\hat{\mu}_2) - \mu = \frac{n}{n+\lambda}\mu - \mu = -\frac{\lambda}{n+\lambda}\mu$$

$$\text{var}\left(\hat{\mu}_2\right) = \text{var}\left(\frac{\sum_i^n y_i}{n+\lambda}\right) = \frac{1}{(n+\lambda)^2}\sum_{i=1}^{n}\text{var}(y_i) = \frac{n}{(n+\lambda)^2}.$$

Note that $\text{var}\left(\hat{\mu}_2\right) < \text{var}(\hat{\mu}_1)$ since $\frac{n}{(n+\lambda)^2} < \frac{1}{n}$ as $\lambda > 0$.

(c) Find a Bayes estimator assuming the $\text{N}(0, 1/\lambda)$ as prior for μ. Compare with your answer in the previous part.

Answer: Using the result of exercise 1, or as shown in the lecture slides, the posterior is also Normal distribution and therefore its mean will be the Bayes estimator. This is equal to

$$\hat{\mu}_3 = \frac{\frac{1}{\lambda}\bar{y}}{\frac{1}{\lambda} + \frac{1}{n}}$$

Multiplying both numerator and denominator with $n\lambda$ gives

$$\hat{\mu}_3 = \frac{\sum_i^n y_i}{n+\lambda}$$

which is the same as $\hat{\mu}_2$.

3. Load the dataset 'diabetes' from the scikit-learn library in Python. Fit a linear regression and a ridge regression model and assess their predictive performance by splitting the data into a training and test dataset.

Answer: See jupyter notebook file Exercise04.ipynb