ST451 Bayesian Machine Learning
Week 4

Exercises

1. Consider the case of Linear Discriminant Analysis with a scalar x with data $(y_i, x_i)_{i=1}^n$, where y_i's are binary random variables and x_i's continuous. Assume that $x_i \sim N(\mu_0, \sigma^2)$ in category c_0 and that $x_i \sim N(\mu_1, \sigma^2)$ in category c_1 and that they are independent. Further assume that each y_i is a Bernoulli random variable with probability of success $p(y \in c_1|x)$ and that the y_i's are independent. Finally, the prior probability $\pi(y \in c_1) = \pi$. Write down the likelihood function and provide the maximum likelihood estimators for π, μ_0, μ_1 and σ^2.

   Answer: The likelihood for $\theta = (\pi, \mu_1, \mu_2, \sigma^2)$ based $(y_i, X_i)_{i=1}^n$ can be written as

$$f(x, y|\theta) = \prod_{i=1}^n \left[\pi N(\mu_1, \sigma^2)\right]^{y_i} \left[(1-\pi)N(\mu_0, \sigma^2)\right]^{1-y_i}$$

   To maximise with respect to π we write the log-likelihood keeping the terms that involve π

$$\log f(x, y|\pi) = c + \sum_{i=1}^n \{y_i \log \pi + (1-y_i)\log(1-\pi)\}$$

   After differentiating the above wrt π, setting equal to 0 and solving the equation we get

$$\hat{\pi} = \frac{1}{n}\sum_{i=1}^n y_i = \frac{n_1}{n} = \frac{n_1}{n_0 + n_1}$$

   To maximise with respect to μ_0 we write the log-likelihood keeping the terms that involve μ_0:

$$\log f(x, y|\mu_1) = c + \sum_{i=1}^n y_i \log N(x_i|\mu_1, \sigma^2) = c - \frac{1}{2}\frac{\sum_{i=1}^n y_i(x_i - \mu_1)^2}{\sigma^2}$$

   After differentiating the above wrt π, setting equal to 0 and solving the equation we get

$$\hat{\mu}_1 = \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n y_i} = \frac{\sum_{i=1}^n y_i x_i}{n_1}$$

   Similarly we obtain

$$\hat{\mu}_0 = \frac{\sum_{i=1}^n (1-y_i) x_i}{\sum_{i=1}^n (1-y_i)} = \frac{\sum_{i=1}^n (1-y_i) x_i}{n_0}$$

   Finally for the common variance σ^2

$$\begin{aligned}
\log f(x, y|\sigma^2) &= c + \sum_{i=1}^n (1-y_i)\log N(x_i|\mu_0, \sigma^2) + \sum_{i=1}^n y_i \log N(x_i|\mu_1, \sigma^2) \\
&= c - (n/2)\log\sigma^2 - \frac{1}{2}\frac{\sum_{i=1}^n (1-y_i)(x_i-\mu_0)^2}{\sigma^2} - \frac{1}{2}\frac{\sum_{i=1}^n y_i(x_i-\mu_1)^2}{\sigma^2} \\
&= c - (n/2)\log\sigma^2 - \frac{1}{2}\frac{\sum_{i=1}^n \left\{(1-y_i)(x_i-\mu_0)^2 + y_i(x_i-\mu_1)^2\right\}}{\sigma^2}
\end{aligned}$$

After differentiating the above wrt σ^2, setting equal to 0 and solving the equation we get

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left\{(1-y_i)(x_i-\hat{\mu}_0)^2 + y_i(x_i-\hat{\mu}_1)^2\right\}$$

2. Let $y = (y_1, \ldots, y_n)$ be a r.s. from a Poisson(λ) and assign the an improper prior to λ such that $\pi(\lambda) \propto \lambda^{-1/2}$. Find the Laplace approximation to the posterior based on the mode and the Hessian matrix of $\pi^*(\lambda|y) = f(y|\lambda)\pi(\lambda)$.

   Answer: We can write

$$\begin{aligned}\log \pi^*(\lambda|y) &= \log\left[f(y|\lambda)\pi(\lambda)\right] \propto \log\left\{\lambda^{\sum y_i}\exp(-n\lambda)\right\} - \frac{1}{2}\log\lambda \\ &= (\sum y_i \log(\lambda) - n\lambda - \frac{1}{2}\log\lambda,\end{aligned}$$

$$\frac{\partial \log \pi^*(\lambda|y)}{\partial\lambda} = \frac{\sum y_i - 1/2}{\lambda} - n$$

   Setting $\frac{\partial \log \pi^*(\lambda|y)}{\partial\lambda} = 0$, gives $\lambda_M = \frac{-1/2+\sum_i y_i}{n}$.
   We also note that

$$\frac{\partial^2 \log \pi^*(\lambda|y)}{\partial\lambda^2} = -\frac{-1/2+\sum_i y_i}{\lambda^2},$$

   which is negative (when evaluated at λ_M) implying the mode is at λ_M. The Hessian is

$$H(\lambda) = \frac{-1/2+\sum_i y_i}{\lambda^2}$$

   The normal approximation to the posterior for λ will have mean λ_M and variance $H(\lambda_M)^{-1}$

3. In the dataset used in the computer class compute the maximum likelihood estimates without using the relevant *sklearn* function but with the *numpy* library only.

   Answer: See jupyter notebook 'CodeExercises.ipynb'

4. The file 'CreditCardFraud.csv' contains a subsample from the following Kaggle competition `https://www.kaggle.com/mlg-ulb/creditcardfraud`. A smaller sample has been taken with more balanced data, as the issue of unbalanced data is not the subject of this week's material. The features 'Amount' and 'Time' were also removed, leaving the 28 features labeled V1-V28. Fit logistic regression models on the data using both the MLE and Bayesian approaches and compare their predictive performance. *Answer:* See jupyter notebook 'CodeExercises.ipynb'

5. The file 'Default.csv' contains the data from the first motivating example in the lecture slides. Fit logistic regression models on the data (both the MLE and Bayesian) as well as the Linear Discriminant Analysis model and compare their predictive performances. *Answer:* See jupyter notebook 'CodeExercises.ipynb'