MLPM Tutorial 3

October 22, 2019

1. Observations $y_0, \ldots, y_{n-1}$ are noisy i.i.d. measurements of an underlying variable x with $p(x) \sim \mathcal{N}(x|0, \sigma_0^2)$ and $p(y_i|x) \sim \mathcal{N}(y_i|x, \sigma^2)$ for $i = 0, \ldots, n-1$. Show that $p(x|y_0, \ldots, y_{n-1})$ is Gaussian with mean

$$\mu = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\bar{y}$$

where $\bar{y} = \frac{1}{n}\sum_{i=0}^{n-1}$ and variance σ_n^2 such that

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}$$

2. Consider the multivariate Gaussian distribution $p(\mathbf{x}) \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ on the vector $\mathbf{x}$ with components $x_1, \ldots, x_n$:

$$p(\mathbf{x}) = \frac{1}{\sqrt{det(2\pi\boldsymbol{\Sigma})}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

Find the mean and covariance of $p(x_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$.

3. Consider a set of data $\mathcal{D} = x_i{}_{i=1}^N$ where each x_i is independently drawn from a Gaussian with known mean μ and unknown variance σ^2. Assume a gamma distribution prior on $\tau = \frac{1}{\sigma^2}$, $p(\tau) = Ga(\tau|a, b)$. Show that the posterior distribution is

$$p(\tau|\mathcal{D}) = Ga\left(\tau|a + \frac{N}{2}, b + \frac{1}{2}\sum_{i=1}^{N}(x_i - \mu)^2\right)$$

4. Consider a 3 class naive Bayes classifier with one binary feature and one Gaussian feature:

$$\begin{aligned} y &\sim Mu(y|\boldsymbol{\pi}, 1) \\ x_1|y = c &\sim Ber(x_1|\theta_c) \\ x_2|y = c &\sim \mathcal{N}(x_2|\mu_c, \sigma_c^2) \end{aligned}$$

Let the parameter vectors be as follows:

$$\begin{aligned} \boldsymbol{\pi} &= (0.5, 0.25, 0.25) \\ \boldsymbol{\theta} &= (0.5, 0.5, 0.5) \\ \boldsymbol{\mu} &= (-1, 0, 1) \\ \boldsymbol{\sigma}^2 &= (1, 1, 1) \end{aligned}$$

Calculate the following quantities:

(a) $p(y|x_1 = 0, x_2 = 0)$
(b) $p(y|x_1 = 0)$
(c) $p(y|x_2 = 0)$

5. Load the height/weight data from `data/heightWeightData.csv`. The first colum is the class label (1 = male, 2 = female), the second column is height (inches), the third weight (lbs). Implement in your favourite programming language the following discriminant analysis models:

(a) $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ where $\boldsymbol{\Sigma}$ is a diagonal matrix.
(b) $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ where $\boldsymbol{\Sigma}$ is a full matrix.
(c) $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are diagonal matrices.
(d) $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are full matrices.

Which one is the best model according to the Bayesian Information Criterion given by

$$BIC = logp(\mathcal{D}|\hat{\boldsymbol{\theta}}_{MLE}) - \frac{d}{2}log(N)$$

where d is the number of free parameters in the model and N is the number of samples? The higher the BIC value, the better the model. Which one is the model with the smallest misclassification rate? Which one is the best model according to BIC or the misclassification rate if you train on 70% of the data and test only on the other 30%?