MLPM Coursework

November 12, 2019

1. A lifetime x of a machine is modelled by an exponential distribution with unknown parameter θ. The likelihood is

$$p(x|\theta) = \text{Expon}(x|\theta) = \theta e^{-\theta x}$$

for $x \geq 0$, $\theta \geq 0$.

   (a) Derive the maximum likelihood estimate $\hat{\theta}_{MLE}$.

   (b) Suppose we observe the lifetime (in years) of 3 different i.i.d. machines, $\mathcal{D} = \{5, 6, 4\}$. What is the MLE given this data?

   (c) Assume that an expert believes θ should have a prior distribution that is also exponential $p(\theta) = \text{Expon}(\theta|\lambda)$. Choose a value $\hat{\lambda}$ for the prior parameter such that $E[\theta] = \frac{1}{3}$.

   (d) Is the exponential prior conjugate to the exponential likelihood?

   (e) What is the posterior, $p(\theta|\mathcal{D}, \hat{\lambda})$?

   (f) What is the MAP estimate $\hat{\theta}_{MAP}$?

   (g) What is the posterior mean estimate $\hat{\theta}_{MEAN}$?

   (h) Why are $\hat{\theta}_{MEAN}$ and $\hat{\theta}_{MLE}$ different? Which is more reasonable in this example?

2. Consider a Naive Bayes model for spam classification with the vocabulary

$$V = \{\text{secret}, \text{offer}, \text{low}, \text{price}, \text{valued}, \text{customer}, \text{today}, \\ \text{dollar}, \text{million}, \text{sports}, \text{is}, \text{for}, \text{play}, \text{healthy}, \text{pizza}\}$$

We have the following example spam messages "million dollar offer", "secret offer today", "secret offer today", "secret is secret", "low price offer" and normal messages "low price for valued customer", "play secrets sports today", "sports is healthy", "low price pizza". Assume the words in a message are independent and identically distributed.

(a) Formally define the model by specifying all needed random variables, parameters, prior, likelihood and posterior distributions.

(b) How many parameters should be learnt from data?

(c) What is the probability of the message "low price today" to be a spam message? Calculate the probability using the maximum likelihood estimates of the required parameters.

(d) Is the assumption about word independence a reasonable one? How can we relax this assumption?

3. Let $\mathbf{x} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$ where $\mathbf{x} \in \mathbb{R}^2$, $\mu = [\mu_1 \quad \mu_2]^T$, $\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$ and ρ is the correlation coefficient.

   (a) What is the joint distribution $p(x_1, x_2)$?

   (b) What is the conditional probability $p(x_2|x_1)$?

   (c) Assume $\sigma_1 = \sigma_2 = 1$. What is $p(x_2|x_1)$ now?

   Simplify your answers as much as possible (e.g. they should not contain any matrix operations).

4. Let $x \in \{0, 1\}$ denote the result of a coin toss ($x = 0$ for tails, $x = 1$ for heads). The coin is potentially biased, so that heads occurs with probability θ_1. Suppose that someone else observes the coin flip and reports to you the outcome, y. But this person is unreliable and only reports the result correctly with probability θ_2. Assuming that θ_2 is independent of x and θ_1 the joint likelihood factorises into

$$p(x, y|\boldsymbol{\theta}) = p(y|x, \theta_2)p(x|\theta_1)$$

   where $\boldsymbol{\theta} = [\theta_1 \quad \theta_2]$.

   (a) Write down the conditional distribution $p(y|x, \theta_2)$ as a 2×2 table in terms of θ_2.

(b) Write down the joint distribution $p(x, y|\boldsymbol{\theta})$ as a 2×2 table in terms of θ_1 and θ_2.

(c) Suppose the following data is observed: $\mathbf{x} = \{1, 1, 0, 1, 1, 0, 0\}$ and $y = \{1, 0, 0, 0, 1, 0, 1\}$. Find the maximum likelihood estimates $\hat{\theta_1}$ and $\hat{\theta_2}$.

(d) What is $p(\mathcal{D}|\hat{\boldsymbol{\theta}}_{MLE}, M_2)$ where M_2 denotes this 2-parameter model?

(e) Now consider a model with 4 parameters, $\boldsymbol{\theta} = [\theta_{0,0} \;\; \theta_{0,1} \;\; \theta_{1,0} \;\; \theta_{1,1}]$, representing $p(x, y|\boldsymbol{\theta}) = \theta_{x,y}$. What is the MLE of $\boldsymbol{\theta}$? What is $p(\mathcal{D}|\hat{\boldsymbol{\theta}}_{MLE}, M_4)$ where M_4 denotes this 4-parameter model?

(f) The Bayesian Information Criterion is defined as

$$\text{BIC}(M, \mathcal{D}) = \log p(\mathcal{D}|\hat{\boldsymbol{\theta}}_{MLE}) - \frac{\text{dof}(M)}{2} \log N$$

where $\text{dof}(M)$ is the number of free parameters (degrees of freedom) in the model and N is the size of the dataset $\mathcal{D}$. Compute the BIC score for both M_2 and M_4 models using log base e. Which model does BIC prefer?

5. Consider fitting a model of the form $p(y|x, \mathbf{w}) = \mathcal{N}(y|w_0 + w_1 x, \sigma^2)$ to the data shown below:

$$\mathtt{x} = [94, 96, 94, 95, 104, 106, 108, 113, 115, 121, 131]$$
$$\mathtt{y} = [0.47, 0.75, 0.83, 0.98, 1.18, 1.29, 1.40, 1.60, 1.75, 1.90, 2.23]$$

For this question you should use your favourite programming language, however you are not allowed to use any specialised machine learning modules. You can use linear algebra and plotting modules, but you should implement everything else yourself. You should clearly state what language and modules you have used and add formatted, preferably syntax highlighted, code snippets for each answer.

(a) Plot the data. Is the proposed linear model a reasonable choice?

(b) Find the ordinary least squares estimate $\hat{\mathbf{w}}_{OLS}$.

(c) Compute an unbiased estimate of σ^2 using

$$\hat{\sigma}^2 = \frac{1}{N-2} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

where $\hat{y}_i = \hat{w}_0 + \hat{w}_1 x_i$ and $\hat{\mathbf{w}}_{OLS} = [\hat{w}_0 \quad \hat{w}_1]$. The denominator is $N - 2$ since we have already estimated 2 parameters from the dataset which reduces the effective sample size by 2.

(d) Now assume the following prior on $\mathbf{w}$:

$$p(\mathbf{w}) = p(w_0)p(w_1)$$

Use an (improper) uniform prior on w_0 and a $\mathcal{N}(0, 1)$ on w_1. Show that this can be written as a Gaussian prior of the form $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w_0}, \mathbf{V_0})$. What are $\mathbf{w}_0$ and $\mathbf{V}_0$?

(e) Derive the posterior $p(\mathbf{w}|\mathcal{D})$ using the prior from above. Plot the distribution (e.g. use a filled contour plot).

(f) Sample lines from the posterior and plot them in the data domain together with the data. Visualise the posterior probability of a line on the plot by changing the colour, thickness or transparency of each line.

Hint: You can generate samples $\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ by first drawing independent samples $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, I)$ and then applying the transformation $\mathbf{w} = \boldsymbol{\mu} + L\mathbf{u}$ where $\Sigma = LL^T$ is the Cholesky decomposition of the desired covariance matrix Σ and I is the identity matrix.

(g) Compute the marginal posterior of the slope, $p(w_1|\mathcal{D}, \hat{\sigma}^2)$, using the data and estimates from above. What is the mean and variance of w_1 according to the marginal posterior?

(h) Compute the conditional posterior of the slope, $p(w_1|w_0 = \hat{w}_0, \mathcal{D}, \hat{\sigma}^2)$, using the data and estimates from above. What is the mean and variance of w_1 according to the conditional posterior?