ELL784/EEL709: Introduction to Machine Learning

Re-Minor Test

April 29, 2016

Maximum marks: 20
Closed book/notes; one two-sided A4 cheat sheet permitted

Please show all steps in your working fully and clearly, except where indicated otherwise. In this question paper, bold symbols indicate vectors. In your answers, please follow a consistent notation to denote vectors (e.g., an underbar).

1. One Sunday morning, the arrival times of six successive north-bound trains at the Hauz Khas Metro station were observed to be as follows: 7:06; 7:13; 7:47; 7:53; 8:01; 8:08. Let us denote by k the time interval (in minutes) between the arrival of two successive trains; we will assume that this follows a *Gaussian distribution*, i.e.,

$$p(k|\lambda,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(k-\lambda)^2}{2\sigma^2}\right\}.$$

(a) The expected value of k under the Gaussian distribution is λ; i.e., in our case, λ corresponds to the expected time interval between successive arrivals. Given a sequence of independent observations of time intervals, $\{k_1, k_2, ..., k_N\}$, write down the likelihood function for this data under the Gaussian distribution, and obtain general expressions for the maximum likelihood estimates for both λ and σ, the standard deviation parameter. Now plug in the Metro data above to get the maximum likelihood estimates for the expected time interval at Hauz Khas station and its standard deviation, based on the five observations. Comment on the meaningfulness of these estimates. **[4]**

(b) Suppose we now treat λ as a random variable and put a prior distribution on it. We may use another Gaussian for this:

$$p(\lambda|\alpha,\beta) = \frac{1}{\sqrt{2\pi}\beta}\exp\left\{-\frac{(\lambda-\alpha)^2}{2\beta^2}\right\}.$$

Here the α and β are 'hyperparameters' to be chosen. Use this prior and the likelihood function from part (a) to obtain the posterior distribution of λ, as a function of the observed data, α, β, and σ (which we assume to be fixed). Derive a general expression for the maximum a posteriori (MAP) estimate for λ. **[3]**

(c) The parameter β above specifies the spread or variance of the prior distribution. Suppose we set $\beta = \sigma/3$. Suppose also that you have heard somewhere that the time interval between successive north-bound trains at Hauz Khas station is about 5 minutes. What would be a reasonable choice for α, in this case? (Look at the MAP expression derived in part (b).) Plug in your choice, along with $\beta = \sigma/3$ and the Metro data given above, to obtain the MAP estimate for the waiting time at Hauz Khas. Comment on the difference between this and the maximum likelihood estimate obtained in part (a). **[3]**

2. (a) In class we saw how the kernel trick can be used to learn non-linear decision boundaries in a given feature space, by implicitly mapping the data to a different space where the boundary

is linear. Suppose K_1 and K_2 are kernels operating on D-dimensional inputs, which implicitly map the inputs to an M-dimensional space. Further, let the feature mappings corresponding to these kernels be $\boldsymbol{\phi}_1 : \mathbb{R}^D \rightarrow \mathbb{R}^M$, $\boldsymbol{\phi}_2 : \mathbb{R}^D \rightarrow \mathbb{R}^M$. Let c be a real-valued constant. Show how $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ can be used to compute the following kernels: (i) $K(\mathbf{x}, \mathbf{x}') = cK_1(\mathbf{x}, \mathbf{x}')$; (ii) $K(\mathbf{x}, \mathbf{x}') = K_1(\mathbf{x}, \mathbf{x}')K_2(\mathbf{x}, \mathbf{x}')$. **[2]**

(b) Suppose we have a two-class data set in 2-D space, generated as follows: positive samples taken from points on the curve $x_1^2 + x_2^2 = 5$, and negative samples taken from points on the curve $x_1^2 + x_2^2 = 10$. The number of samples in both classes are equal. Show visually the kind of decision boundary that would be obtained by training an SVM with a (i) linear kernel; (ii) polynomial kernel of order 2; (iii) RBF kernel. **[2]**

(c) You are given a small data set with just 4 points in 2-D space. Two positive examples, with coordinates $(1, 4)$ and $(2, 3)$; and two negative examples, with coordinates $(4, 5)$ and $(5, 6)$. Find the weight vector $\mathbf{w}$ (including the bias term w_0) corresponding to the maximum-margin decision boundary learnt by an SVM on this data set. Give justification/derivation for your answer. Also draw a plot showing the data points (with support vectors circled) and the decision boundary learnt. **[3]**

3. Given 3 data points in 2-D space, $(1, 1)$, $(2, 2)$, and $(3, 3)$, answer the following:

(a) What is the first principal component? Write down a *unit vector* giving its direction. **[1]**

(b) If we want to project the original data points into 1-D space via PCA, what will be the variance of the projected data? **[1]**

(c) For the projected data in part (b), if we now represent them in the original 2-D space, what is the reconstruction error? **[1]**