Mock midterm for Pattern Classification and Machine Learning, 2015

Room INJ218, Thursday Nov. 12 from 14:15 to 16:00

Teacher: Mohammad Emtiyaz Khan

A few important informations:

- The exam is worth a total of 30 marks.
- You are not allowed to enter after 14:30 and leave before 15:00.
- No electronic devices are allowed except a calculator. Make sure that your calculator is only a calculator and cannot be used for any other purpose.
- Please leave your other belongings in front of the room (or at the back).
- You are not allowed to talk to others
- The mock midterm is not graded or corrected by the teaching team.
- Solutions will be available in December.
- There are extra pages at the end of the exam. Ask us if you need more pages.
- := means "defined as".
- For derivations, clearly explain your derivation step by step. In the final exam you will be marked for steps as well as for the end result.
- We will denote the output data vector by $\mathbf{y}$ which is a vector that contains all y_n, and the feature matrix by $\mathbf{X}$ which is a matrix containing features $\mathbf{x}_n^T$ as rows. Also, $\widetilde{\mathbf{x}}_n = [1, \mathbf{x}_n^T]^T$.

1 Kernels [5 marks in total]

Consider the following function over feature vectors $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^D$:

$$K(\mathbf{x}_i, \mathbf{x}_j) = (1 + a\,\mathbf{x}_i^T \mathbf{x}_j)^2, \quad a \in \mathbb{R}, \quad a > 0 \tag{1}$$

(A) **[2 marks]** Name two properties the function $K(\mathbf{x}_i, \mathbf{x}_j)$ must have for it to be a kernel.

(B) **[3 marks]** Show that the function $K(\mathbf{x}_i, \mathbf{x}_j)$ is a kernel.

Hint: The proof might be easier if you expand $(1 + a\mathbf{x}_i^T \mathbf{x}_j)^2$ and use the fact that the functions $\mathbf{x}_i^T \mathbf{x}_j$ and $(\mathbf{x}_i^T \mathbf{x}_j)^2$ are kernels and follow the two properties.

2 Multiple-output regression [5 marks in total]

Suppose we have N regression training-pairs, but instead of one output for each input vector $\mathbf{x}_n \in \mathbb{R}^D$, we now have multiple outputs $\mathbf{y}_n = [y_{n1}, y_{n2}, \ldots, y_{nK}]^T \in \mathbb{R}^K$. For each output y_{nk}, we wish to fit a separate linear model:

$$y_{nk} \approx f_k(\mathbf{x}_n) = \beta_{k1}x_{n1} + \beta_{k2}x_{n2} + \ldots + \beta_{kD}x_{nD} = \boldsymbol{\beta}_k^T \mathbf{x}_n \tag{2}$$

where $\boldsymbol{\beta}_k$ is the vector of β_{kd} for $d = 1, 2, \ldots, D$. Note that there is no bias term.

Our goal is to estimate $\boldsymbol{\beta} = [\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_K^T]^T$ for which we choose to minimize the following cost function:

$$\mathcal{L}(\boldsymbol{\beta}) := \sum_{k=1}^{K} \sum_{n=1}^{N} \frac{1}{2\sigma_k^2} \left(y_{nk} - \boldsymbol{\beta}_k^T \mathbf{x}_n\right)^2 + \frac{1}{2\sigma_0^2} \sum_{k=1}^{K} \sum_{d=1}^{D} \beta_{kd}^2 \tag{3}$$

where $\sigma_k > 0$ are known real-valued scalars for $k = 0, 1, \ldots, K$. We denote the set of all σ_k by $\boldsymbol{\sigma}$.

(A) [**1 mark**] Derive the normal equation for $\boldsymbol{\beta}_k^*$ that minimizes $\mathcal{L}$.

(B) [**2 marks**] Discuss the conditions under which the minimum $\boldsymbol{\beta}_k^*$ is unique. Assuming the conditions hold, write the expression for the unique solution.

(C) [**2 marks**] Let $\boldsymbol{\beta}^*$ be the vector of all $\boldsymbol{\beta}_k^*$. Derive a probabilistic model under which the solution $\boldsymbol{\beta}^*$ is the maximum-a-posteriori (MAP) estimate. You must give expressions for the likelihood $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\sigma})$ and the prior $p(\boldsymbol{\beta}|\boldsymbol{\sigma})$.

3 Mixture of Linear Regression [10 marks in total]

In Project-I, you worked on a regression dataset with two or more distinct clusters. For such datasets, a mixture of linear regression models is preferred over just one linear regression model.

Consider a regression dataset with N pairs $\{y_n, \mathbf{x}_n\}$. Similar to Gaussian mixture-model (GMM), let $r_n \in \{1, 2, \ldots, K\}$ index the mixture component. Distribution of the output y_n under the k'th linear model is defined as follows:

$$p(y_n|\mathbf{x}_n, r_n = k, \boldsymbol{\beta}) := \mathcal{N}(y_n|\boldsymbol{\beta}_k^T \widetilde{\mathbf{x}}_n, 1) \tag{4}$$

Here, $\boldsymbol{\beta}_k$ is the regression parameter vector for the k'th model with $\boldsymbol{\beta}$ being a vector containing all $\boldsymbol{\beta}_k$. Also, $\widetilde{\mathbf{x}}_n = [1, \mathbf{x}_n^T]^T$.

(A) **[2 marks]** Define $\mathbf{r}_n$ to be a binary vector of length K such that all the entries are 0 except a k'th entry i.e. $r_{nk} = 1$, implying that $\mathbf{x}_n$ is assigned to the k'th mixture. Rewrite the likelihood $p(y_n|\mathbf{x}_n, \boldsymbol{\beta}, \mathbf{r}_n)$ in terms of r_{nk}.

(B) **[1 mark]** Write the expression for the joint distribution $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \mathbf{r})$ where $\mathbf{r}$ is the set of all $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N$.

(C) **[3 marks]** Assume that r_n follows a multinomial distribution $p(r_n = k|\boldsymbol{\pi}) = \pi_k$, with $\boldsymbol{\pi} = [\pi_1, \pi_2, \ldots, \pi_K]$. Derive the marginal distribution $p(y_n|\mathbf{x}_n, \boldsymbol{\beta}, \boldsymbol{\pi})$ obtained after marginalizing r_n out.

(D) **[2 marks]** Write the expression for the maximum likelihood estimator $\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\pi}) := -\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\pi})$ in terms of data $\mathbf{y}$ and $\mathbf{X}$, and parameters $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$.

(E) **[2 marks]** Is $\mathcal{L}$ jointly-convex with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$? Is the model identifiable? Prove your answers.

4 Multi-class classification [5 marks in total]

Suppose we have a classification dataset with N pairs $\{y_n, \mathbf{x}_n\}$ but now y_n is a categorical variable, i.e. $y_n \in \{1, 2, \ldots, K\}$ where K is the number of classes. We wish to fit a linear model and in the similar spirit to logistic regression, we will use a multinomial logit distribution to map linear inputs to a categorical output.

We will define $\eta_{nk} = \widetilde{\mathbf{x}}_n^T \boldsymbol{\beta}_k$ for all $k = 1, 2, \ldots, K-1$ and then compute the probability of output,

$$p(y_n = k | \mathbf{x}_n, \boldsymbol{\beta}) = \frac{e^{\eta_{nk}}}{\sum_{j=1}^{K} e^{\eta_{nj}}} \tag{5}$$

For identifiability reasons, we set $\eta_{nK} = 0$, therefore $\boldsymbol{\beta}_K = 0$ and we need to estimate $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_{K-1}$.

Similar to logistic regression, we will assume that each y_n is i.i.d. i.e.

$$p(\mathbf{y} | \mathbf{X}, \boldsymbol{\beta}) = \prod_{n=1}^{N} p(y_n | \mathbf{x}_n, \boldsymbol{\beta}) \tag{6}$$

Following the derivation of logistic regression,

(A) **[2 marks]** Derive the log-likelihood for this model.

(B) **[2 marks]** Derive the gradient with respect to $\boldsymbol{\beta}_k$.

(C) **[1 marks]** Show that the negative of the log-likelihood is convex.

5 Proportional Hazard Model [5 marks in total]

We have a regression dataset with N pairs $\{y_n, \mathbf{x}_n\}$ where the output is an ordered output i.e. $y_n \in \{1, 2, 3, 4, \ldots, K\}$ (as opposed to an un-ordered output in the standard multi-class classification). We wish to fit a linear model.

In the proportional hazard model, we use the following probability distribution,

$$p(y_n = k|\mathbf{x}_n, \boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{\exp(\eta_{nk})}{\sum_{j=1}^{K} \exp(\eta_{nj})}, \text{ where } \eta_{nk} = \theta_k + \boldsymbol{\beta}^T \mathbf{x}_n, \forall k \tag{7}$$

Here, $\theta_k \in \mathbb{R}$ and are ordered, i.e. $\theta_1 > \theta_2 > \ldots > \theta_K$. We will denote the vector of all θ_k by $\boldsymbol{\theta}$. Similar to a standard regression model, we assume that all pairs $\{y_n, \mathbf{x}_n\}$ are i.i.d.

Answer the following questions. Clearly show all steps of your derivations.

(A) [**2 marks**] Is $p(y_n|\mathbf{x}_n, \boldsymbol{\beta}, \boldsymbol{\theta})$ a valid distribution? Prove your answer.

Hint: You need to prove two properties to be able to show this.

(B) [**2 marks**] Derive the log-likelihood for this model.

(C) [**1 marks**] Show that the negative of the log-likelihood is convex w.r.t. all θ_k and $\boldsymbol{\beta}$.

Notes

Notes