Mock midterm for Pattern Classification and Machine Learning, 2015

Solutions

1 Kernels [5 marks in total]

Answer:
(A) To prove that the function is Kernel, we need two properties: symmetry and positive-semi-definiteness (p.s.d.).

(B) The function is symmetric.

$$K(\mathbf{x}_i, \mathbf{x}_j) = (1 + a\mathbf{x}_i^T\mathbf{x}_j)^2 = (1 + a\mathbf{x}_j^T\mathbf{x}_i)^2 = K(\mathbf{x}_j, \mathbf{x}_i) \tag{1}$$

To prove p.s.d. property, we need that the product $\mathbf{t}^T\mathbf{K}\mathbf{t} \geq 0$ for all non-zero vectors $\mathbf{t}$ ($\mathbf{K}$ is the $N \times N$ matrix formed with N features). To prove this, we rewrite the product as follows,

$$\mathbf{t}^T\mathbf{K}\mathbf{t} = \sum_i \sum_j K_{ij} t_i t_j = \sum_i \sum_j (1 + a\mathbf{x}_i^T\mathbf{x}_j)^2 t_i t_j \tag{2}$$

$$= \sum_i \sum_j (1 + a\mathbf{x}_i^T\mathbf{x}_j)(1 + a\mathbf{x}_i^T\mathbf{x}_j) t_i t_j \tag{3}$$

$$= \sum_i \sum_j \left[t_i t_j + 2a\mathbf{x}_i^T\mathbf{x}_j t_i t_j + a^2(\mathbf{x}_i^T\mathbf{x}_j)^2 t_i t_j \right] \tag{4}$$

$$= \left(\sum_i t_i \right)^2 + 2a \sum_i \sum_j \mathbf{x}_i^T\mathbf{x}_j t_i t_j + \sum_i \sum_j a^2(\mathbf{x}_i^T\mathbf{x}_j)^2 t_i t_j \tag{5}$$

The first term is positive and the next two terms are positive since the two functions are kernels as well.

2 Multiple-output regression [5 marks in total]

Answer:
(A) The cost function can be split into K summands $\mathcal{L}(\boldsymbol{\beta}) = \sum_{k=1}^{K} \mathcal{L}(\boldsymbol{\beta}_k)$. For each $\boldsymbol{\beta}_k$ we can then write:

$$\mathcal{L}(\boldsymbol{\beta}_k) := \sum_{n=1}^{N} \frac{1}{2\sigma_k^2} \left(y_{nk} - \boldsymbol{\beta}_k^T\mathbf{x}_n \right)^2 + \frac{1}{2\sigma_0^2} \sum_{j=1}^{D} \beta_{kj}^2 \tag{6}$$

$$= \frac{1}{2\sigma_k^2} (\mathbf{y}_k - \mathbf{X}\boldsymbol{\beta}_k)^T (\mathbf{y}_k - \mathbf{X}\boldsymbol{\beta}_k) + \frac{1}{2\sigma_0^2} \boldsymbol{\beta}_k^T \boldsymbol{\beta}_k \tag{7}$$

Taking the derivative with respect to $\boldsymbol{\beta}_k$ and setting it to zero yields the normal equation:

$$\frac{1}{\sigma_k^2} \mathbf{X}^T(\mathbf{X}\boldsymbol{\beta}_k^* - \mathbf{y}) + \frac{1}{\sigma_0^2} \boldsymbol{\beta}_k^* = 0 \tag{8}$$

This is a slight modification of Ridge regression. Note that minimizing each of those summands individually also minimizes the total $\mathcal{L}(\boldsymbol{\beta})$.

(B) There are no conditions. Even if $\mathbf{X}$ is not full column rank, the solution will exist. The regularization term adds a $\frac{1}{\sigma_0^2}\mathbf{I}_D$ (see below), which makes it such that the expression in the brackets is invertible.

$$\left(\frac{1}{\sigma_k^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\sigma_0^2}\mathbf{I}_D\right)\boldsymbol{\beta}_k^* = \frac{1}{\sigma_k^2}\mathbf{X}^T\mathbf{y} \tag{9}$$

$$\Rightarrow \boldsymbol{\beta}_k^* = \left(\frac{1}{\sigma_k^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\sigma_0^2}\mathbf{I}_D\right)^{-1}\frac{1}{\sigma_k^2}\mathbf{X}^T\mathbf{y} \tag{10}$$

(C) You are asked to derive a probabilistic model under which $\boldsymbol{\beta}^*$ is the maximum a posteriori estimate, i.e. $\arg\max_{\boldsymbol{\beta}} \text{Posterior} = \boldsymbol{\beta}^*$. However, since "Posterior probability $\propto$ Likelihood $\times$ Prior probability", the questions asks specifically for the prior and the likelyhood only. Knowing that the maximization over a Gaussian is equivalent to minimizing the mean square error, one can check that $\boldsymbol{\beta}_{MAP}^* = \arg\max_{\boldsymbol{\beta}} p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta})p(\boldsymbol{\beta})$ is equivalent to the above cost minimization $\boldsymbol{\beta}_{normal}^* = \arg\min_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta})$ if:

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\mathcal{N}(y_{nk}|\boldsymbol{\beta}_k^T\mathbf{x}_n, \sigma_k^2) \tag{11}$$

$$p(\boldsymbol{\beta}) = \prod_{k=1}^{K}\mathcal{N}(\boldsymbol{\beta}_k|0, \sigma_0^2\mathbf{I}_D) \tag{12}$$

3 Mixture of Linear Regression [15 marks in total]

Answer:
(A) Likelihood: $p(y_n|\mathbf{x}_n, \boldsymbol{\beta}, \mathbf{r}_n) = \prod_{k=1}^{K}\left[\mathcal{N}(y_n|\boldsymbol{\beta}_k^T\widetilde{\mathbf{x}}_n, \sigma^2)\right]^{r_{nk}}$

(B) Joint likelihood: $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \mathbf{r}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\left[\mathcal{N}(y_n|\boldsymbol{\beta}_k^T\widetilde{\mathbf{x}}_n, \sigma^2)\right]^{r_{nk}}$

(C) Write the joint, then the conditional, and plug in.

$$\begin{aligned} p(y_n|\mathbf{x}_n, \boldsymbol{\beta}, \boldsymbol{\pi}) &= \sum_{k=1}^{K} p(y_n, r_n = k|\mathbf{x}_n, \boldsymbol{\beta}, \boldsymbol{\pi}) = \sum_{k=1}^{K} p(y_n|r_n = k, \mathbf{x}_n, \boldsymbol{\beta}, \boldsymbol{\pi})\pi_k \\ &= \sum_{k=1}^{K}\mathcal{N}(y_n|\boldsymbol{\beta}_k^T\widetilde{\mathbf{x}}_n, \sigma^2)\pi_k \end{aligned} \tag{13}$$

(D)

$$-\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\pi}) = -\log \prod_{n=1}^{N} \sum_{k=1}^{K} \mathcal{N}(y_n|\boldsymbol{\beta}_k^T \widetilde{\mathbf{x}}_n, \sigma^2)\pi_k \tag{14}$$

$$= -\sum_{n=1}^{N} \log \sum_{k=1}^{K} \mathcal{N}(y_n|\boldsymbol{\beta}_k^T \widetilde{\mathbf{x}}_n, \sigma^2)\pi_k \tag{15}$$

(E) A model is identifieable iff $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 \rightarrow P_{\boldsymbol{\theta}_1} = P_{\boldsymbol{\theta}_2}$, i.e. the relationship of the parameters to the model is one to one. The given model is not identifieable for two reasons:

- By permutation of labels.
- Imagine two models with $\boldsymbol{\beta}_k^T \widetilde{\mathbf{x}}_n = 0 \ \forall(k, n)$ which are identical except for two different sets of $\boldsymbol{\pi}$: $\boldsymbol{\pi}^*$ and $\hat{\boldsymbol{\pi}}$.

  Then $$\sum_{k=1}^{K} \mathcal{N}(y_n|\boldsymbol{\beta}_k^T \widetilde{\mathbf{x}}_n, \sigma^2)\pi_k^* = \sum_{k=1}^{K} \mathcal{N}(y_n|\boldsymbol{\beta}_k^T \widetilde{\mathbf{x}}_n, \sigma^2)\hat{\pi}_k \tag{16}$$

  imples that $$\sum_{k=1}^{K} \mathcal{N}(y_n|0, \sigma^2)\pi_k^* = \sum_{k=1}^{K} \mathcal{N}(y_n|0, \sigma^2)\hat{\pi}_k \tag{17}$$

  imples that $$\mathcal{N}(y_n|0, \sigma^2) \sum_{k=1}^{K} \pi_k^* = \mathcal{N}(y_n|0, \sigma^2) \sum_{k=1}^{K} \hat{\pi}_k \tag{18}$$

  imples that $$\mathcal{N}(y_n|0, \sigma^2) = \mathcal{N}(y_n|0, \sigma^2) \tag{19}$$

  And the last statement is of course true. Since we started with a set of different $\boldsymbol{\pi}$, this should *never* happen if the model was identifieable.

The model is also *not convex*, since a sum of Gaussians is not convex to begin with.

4 Multi-class classification [5 marks in total]

Answer:
(A) Define the vector $\widetilde{\mathbf{y}}_n$ such that $\widetilde{y}_{nk} = 1$ when $y_n = k$ and rest of entries of $\widetilde{\mathbf{y}}_n$

are zero.

$$\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) = \log \prod_{n=1}^{N} p(y_n|\mathbf{x}_n, \boldsymbol{\beta}) \tag{20}$$

$$= \log \prod_{n:y_n=1} p(y_n = 1|\mathbf{x}_n, \boldsymbol{\beta}) \prod_{n:y_n=2} p(y_n = 2|\mathbf{x}_n, \boldsymbol{\beta}) \dots \prod_{n:y_n=K} p(y_n = K|\mathbf{x}_n, \boldsymbol{\beta}) \tag{21}$$

$$= \log \prod_{k=1}^{K} \prod_{n=1}^{N} [p(y_n = k|\mathbf{x}_n, \boldsymbol{\beta})]^{\widetilde{y}_{nk}} \tag{22}$$

$$= \sum_{k=1}^{K} \sum_{n=1}^{N} \widetilde{y}_{nk} \log p(y_n = k|\mathbf{x}_n, \boldsymbol{\beta}) \tag{23}$$

$$= \sum_{k=1}^{K} \sum_{n=1}^{N} \widetilde{y}_{nk} \left[\eta_{nk} - \log \sum_{j=1}^{K} \exp(\eta_{nj}) \right] \tag{24}$$

$$= \sum_{k=1}^{K} \sum_{n=1}^{N} \widetilde{y}_{nk} \left[\boldsymbol{\beta}_k^T \widetilde{\mathbf{x}}_n - \log \sum_{j=1}^{K} \exp(\boldsymbol{\beta}_j^T \widetilde{\mathbf{x}}_n) \right] \tag{25}$$

$$= \sum_{k=1}^{K} \sum_{n=1}^{N} \widetilde{y}_{nk} \boldsymbol{\beta}_k^T \widetilde{\mathbf{x}}_n - \sum_{n=1}^{N} \log \sum_{j=1}^{K} \exp(\boldsymbol{\beta}_j^T \widetilde{\mathbf{x}}_n) \tag{26}$$

Last step is obtained since $\sum_k y_{nk} = 1$. Notice the similarity to logistic regression.

Note: You get full marks even if you skip the last step. The last step is useful for the next part.

(B) Taking the derivative with respect to $\boldsymbol{\beta}_k$, we get,

$$\sum_{n=1}^{N} \widetilde{y}_{nk} \widetilde{\mathbf{x}}_n - \sum_{n=1}^{N} \frac{\exp(\boldsymbol{\beta}_k^T \widetilde{\mathbf{x}}_n)}{\sum_{j=1}^{K} \exp(\boldsymbol{\beta}_j^T \widetilde{\mathbf{x}}_n)} \widetilde{\mathbf{x}}_n \tag{27}$$

Note: You will get full marks if you write the above expression. Note the similarity with the logistic regression. We can write the normal equation in the same form as logistic regression.

$$\widetilde{\mathbf{X}}^T [\widetilde{\mathbf{y}}_k - \mathcal{S}(\widetilde{\mathbf{X}} \beta_k)] = 0 \tag{28}$$

where $\widetilde{\mathbf{y}}_k$ is the vector containing $\widetilde{y}_{nk}$ for all n, $\mathcal{S}(\boldsymbol{\eta})_j = \frac{\exp(\eta_j)}{\sum_{k=1}^{K} \exp(\eta_k)}$ is the softmax function which is an extension of the σ function to multi-class.

(C) Negative of the log-likelihood is the following:

$$-\sum_{k=1}^{K} \sum_{n=1}^{N} \widetilde{y}_{nk} \boldsymbol{\beta}_k^T \widetilde{\mathbf{x}}_n + \sum_{n=1}^{N} \log \sum_{j=1}^{K} \exp(\boldsymbol{\beta}_j^T \widetilde{\mathbf{x}}_n) \tag{29}$$

Since we know that sum of convex functions is convex, we can ignore the sum over n and we only need to prove that the following is convex:

$$-\sum_{k=1}^{K} \widetilde{y}_{nk} \boldsymbol{\beta}_k^T \widetilde{\mathbf{x}}_n + \log \sum_{j=1}^{K} \exp(\boldsymbol{\beta}_j^T \widetilde{\mathbf{x}}_n) \tag{30}$$

The first term is linear in $\boldsymbol{\beta}_k$, so we only need to prove that the second term is convex which will be true if we prove that log-sum-exp $\log(e^{t_1} + e^{t_2} + \ldots + e^{t_K})$ is a convex function. This was given as an exercise in the online questions so you might know how to prove this. A straightforward way to prove this is to show that the second derivative is positive-definite, although there are other proofs.

Note: We don't expect you to prove convexity of log-sum-exp, rather you should know it as a fact that it is true, since this was covered in class.

5 Proportional Hazard Model [5 marks in total]

Answer:

(A) Show that the probabilities are positive and that they sum to 1. The first property can be seen since $\exp(x) \geq 0 \ \forall x \in \mathbb{R}$. For the second: $\sum_{k=1}^{K} \frac{\exp(\eta_{nk})}{\sum_{j=1}^{K} \exp(\eta_{nj})} = \frac{\sum_{k=1}^{K} \exp(\eta_{nk})}{\sum_{j=1}^{K} \exp(\eta_{nj})} = 1$. Therefore this is a valid probability distribution.

(B) The likelihood is given by the product of the $p(y_n = k|\mathbf{x}_n, \boldsymbol{\beta}, \boldsymbol{\theta})$. Define the vector $\tilde{\mathbf{y}}_n$ such that $y_{nk} = 1$ when $y_n = k$ and the rest of the entries are zero. Taking the log we get:

$$\log \prod_{n=1}^{N} \prod_{k=1}^{K} p(y_n = k|\mathbf{x}_n, \boldsymbol{\beta}, \boldsymbol{\theta})^{\widetilde{y}_{nk}} = \log \prod_{n=1}^{N} \prod_{k=1}^{K} \frac{\exp(\eta_{nk})^{\widetilde{y}_{nk}}}{(\sum_{j=1}^{K} \exp(\eta_{nj}))^{\widetilde{y}_{nk}}} \tag{31}$$

$$= \log \prod_{n=1}^{N} \left[\frac{1}{\sum_{j=1}^{K} \exp(\eta_{nj})} \prod_{k=1}^{K} \exp(\eta_{nk})^{\widetilde{y}_{nk}} \right] \tag{32}$$

$$= \sum_{n=1}^{N} \left[-\log \sum_{j=1}^{K} \exp(\eta_{nj}) + \sum_{k=1}^{K} \widetilde{y}_{nk} \eta_{nk} \right] \tag{33}$$

(C) Since we know that sum of convex functions is convex, it is ok to ignore the sum over n and prove that the following is convex:

$$-\sum_{k=1}^{K} (\theta_k + \boldsymbol{\beta}_k^T \widetilde{\mathbf{x}}_n) + \log \sum_{k=1}^{K} \exp(\theta_k + \boldsymbol{\beta}_k^T \widetilde{\mathbf{x}}_n) \tag{34}$$

First term is linear, so we only need to prove that the second term is convex which will be true if we prove that log-sum-exp $\log(e^{t_1} + e^{t_2} + \ldots + e^{t_K})$ is a convex function. Again, you do not need to prove this, but simply state the fact that it is convex, since this was covered in class.