Midterm Exam

Time: 1:10pm-3:00pm
No discussion is allowed.
You may refer to any related materials.
Use the mathematical notation of PRML as possible as you can.
Gaussian identities:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \tag{2.113}$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \tag{2.114}$$
$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathsf{T}}) \tag{2.115}$$
$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathsf{T}}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) \tag{2.116}$$
$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^{\mathsf{T}}\mathbf{L}\mathbf{A})^{-1} \tag{2.117}$$

---

1. **(20 points)** Bayesian linear regression, posterior → prior

Consider a linear basis function model with the likelihood

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1})$$

and the prior $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$, and suppose that we have already observed N data points, so that the posterior distribution over $\mathbf{w}$ is given by $p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$, where $\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}_N^{\mathsf{T}}\mathbf{t})$ and $\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}_N^{\mathsf{T}}\boldsymbol{\Phi}_N$. The matrix $\boldsymbol{\Phi}_N$ has N rows, each of which is a row vector $\boldsymbol{\phi}(\mathbf{x}_n)^{\mathsf{T}}$. The posterior can be regarded as the prior for the next observation.

(1) Consider an additional data point $(\mathbf{x}_{N+1}, t_{N+1})$, and apply (2.113), (2.114), (2.116). Write down $\mathbf{x}, \mathbf{y}, \mathbf{A}, \mathbf{b}, \mathbf{L}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}$ in (2.113), (2.114), (2.116), in terms of the corresponding variables, means, and (co)variances of the prior and the likelihood. For example, $\mathbf{x} \equiv \mathbf{w}$, $\mathbf{b} \equiv \mathbf{0}$.

(2) Show that the resulting posterior distribution is also given by

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_{N+1}, \mathbf{S}_{N+1}),$$

where $\mathbf{m}_{N+1} = \mathbf{S}_{N+1}(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}_{N+1}^{\mathsf{T}}\mathbf{t})$ and $\mathbf{S}_{N+1}^{-1} = \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}_{N+1}^{\mathsf{T}}\boldsymbol{\Phi}_{N+1}$.

Solution:

(1) By (2.113) and (2.114):

$\mathbf{x} \equiv \mathbf{w}$, $\mathbf{b} \equiv \mathbf{0}$, $\boldsymbol{\mu} \equiv \mathbf{m}_N$, $\boldsymbol{\Lambda}^{-1} \equiv \mathbf{S}_N$, $\mathbf{y} \equiv t_{N+1}$, $\mathbf{A} \equiv \boldsymbol{\phi}(\mathbf{x}_{N+1})^\mathsf{T}$, $\mathbf{L}^{-1} \equiv \beta^{-1}$.

(2) By (2.116) and (2.117):

$\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A} \Rightarrow$

$\mathbf{S}_{N+1}^{-1} = \mathbf{S}_N^{-1} + \beta\boldsymbol{\phi}(\mathbf{x}_{N+1})\boldsymbol{\phi}(\mathbf{x}_{N+1})^\mathsf{T} = \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}_N^\mathsf{T}\boldsymbol{\Phi}_N + \beta\boldsymbol{\phi}(\mathbf{x}_{N+1})\boldsymbol{\phi}(\mathbf{x}_{N+1})^\mathsf{T} = \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}_{N+1}^\mathsf{T}\boldsymbol{\Phi}_{N+1}$.

$\boldsymbol{\Sigma}\{\mathbf{A}^\mathsf{T}\mathbf{L}(\mathbf{y}-\mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\} \Rightarrow$

$\mathbf{m}_{N+1} = \mathbf{S}_{N+1}(\boldsymbol{\phi}(\mathbf{x}_{N+1})\beta t_{N+1} + \mathbf{S}_N^{-1}\mathbf{m}_N) = \mathbf{S}_{N+1}(\boldsymbol{\phi}(\mathbf{x}_{N+1})\beta t_{N+1} + \mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}_N^\mathsf{T}\mathbf{t}) = \mathbf{S}_{N+1}(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}_{N+1}^\mathsf{T}\mathbf{t})$.

2. **(20 points)** Multiclass logistic regression

Consider the posterior probabilities of K classes given by the softmax functions

$$p(\mathcal{C}_k|\boldsymbol{\phi}) = y_k(\boldsymbol{\phi}) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

where the 'activations' a_k are given by

$$a_k = \mathbf{w}_k^\mathsf{T}\boldsymbol{\phi}\,.$$

(1) Show that the derivatives of the softmax are given by

$$\frac{\partial y_k}{\partial a_j} = y_k(I_{kj} - y_j)$$

where I_{kj} are the elements of the identity matrix.

(2) Show that the gradients of the cross-entropy error function (negative logarithm of the likelihood function) are given by

$$\nabla_{\mathbf{w}_j} - \ln p(\mathbf{T}|\mathbf{w}_1, \ldots, \mathbf{w}_K) = \sum_{n=1}^{N}(y_{nj} - t_{nj})\boldsymbol{\phi}_n\,,$$

where $\mathbf{T}$ is an $N \times K$ matrix of target variables. The nth row of $\mathbf{T}$ is the target vector $\mathbf{t}_n$, which is a binary target vector of length K that uses the 1-of-K coding scheme. The matrix $\mathbf{T}$ has elements $t_{nj} = I_{jk}$ if pattern n is from class $\mathcal{C}_k$.

Solution:

*For details, see the lecture notes of week 7, pages 4 and 5.

(1)

$$\frac{\partial y_k}{\partial a_k} = \frac{\exp(a_k)}{\sum_i \exp(a_i)} - \left(\frac{\exp(a_k)}{\sum_i \exp(a_i)}\right)^2 = y_k(1 - y_k)\,,$$

$$\frac{\partial y_k}{\partial a_j} = -\frac{\exp(a_k)\exp(a_j)}{(\sum_i \exp(a_i))^2} = -y_k y_j\,, \text{ for } j \neq k\,.$$

Therefore,

$$\frac{\partial y_k}{\partial a_j} = y_k(I_{kj} - y_j)\,.$$

(2) Let $E \equiv -\ln p(\mathbf{T}|\mathbf{w}_1, \ldots, \mathbf{w}_K) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_{nk}$. We obtain

$$\frac{\partial E}{\partial y_{nk}} = -\frac{t_{nk}}{y_{nk}}\,.$$

Owing to the 1-of-K coding scheme, we have $\sum_k t_{nk} = 1$. By the chain rule:

$$\frac{\partial E}{\partial a_{nj}} = \sum_{k=1}^{K} \frac{\partial E}{\partial y_{nk}} \frac{\partial y_{nk}}{\partial a_{nj}} = -\sum_{k=1}^{K} \frac{t_{nk}}{y_{nk}} y_{nk}(I_{kj} - y_{nj}) = y_{nj} - t_{nj}\,.$$

Again, by the chain rule:

$$\nabla_{\mathbf{w}_j} E = \sum_{n=1}^{N} \frac{\partial E}{\partial a_{nj}} \left(\nabla_{\mathbf{w}_j} a_{nj}\right) = \sum_{n=1}^{N} (y_{nj} - t_{nj})\boldsymbol{\phi}_n\,.$$

3. **(20 points)** Generative classification model and maximum likelihood

Consider a generative classification model for K classes defined by prior class probabilities $p(\mathcal{C}_k) = \pi_k$ and general class-conditional densities $p(\boldsymbol{\phi}|\mathcal{C}_k)$ where $\boldsymbol{\phi}$ is the input feature vector. Suppose we are given a training data set $\{\boldsymbol{\phi}_n, \mathbf{t}_n\}$ where $n = 1, \ldots, N$, and $\mathbf{t}_n$ is a binary target vector of length K that uses the 1-of-K coding scheme, so that it has components $t_{nj} = I_{jk}$ if pattern n is from class $\mathcal{C}_k$. Assuming that the data points are drawn independently from this model, show that the maximum-likelihood solution for the prior probabilities is given by

$$\pi_k = \frac{N_k}{N}$$

where N_k is the number of data points assigned to class $\mathcal{C}_k$.

Solution:

PRML Exercise 4.9. The solution is available on the book web site.

The log-likelihood is given by

$$\ln p(\{\boldsymbol{\phi}_n, \mathbf{t}_n\}|\{\pi_k\}) = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\{\ln p(\boldsymbol{\phi}_n|\mathcal{C}_k) + \ln \pi_k\}\,.$$

We need to satisfy the constraint $\sum_k \pi_k = 1$. Consequently, we maximize

$$\ln p(\{\boldsymbol{\phi}_n, \mathbf{t}_n\}|\{\pi_k\}) + \lambda \left(\sum_k \pi_k - 1 \right) .$$

Setting the derivative with respect to π_k equal to zero, we obtain

$$\sum_{n=1}^{N} \frac{t_{nk}}{\pi_k} + \lambda = 0 \,,$$

and therefore

$$-\pi_k \lambda = \sum_{n=1}^{N} t_{nk} = N_k \,.$$

Summing both sides over k we have $\lambda = -N$, and hence $\pi_k = N_k/N$.

4. **(30 points)** Kernelizing Fisher's linear discriminant for two classes.

   Consider the Fisher criterion in the form

   $$J(\mathbf{w}) = \frac{\mathbf{w}^\mathsf{T} \mathbf{S}_\mathrm{B} \mathbf{w}}{\mathbf{w}^\mathsf{T} \mathbf{S}_\mathrm{W} \mathbf{w}} \,,$$

   as shown in Eq. (4.26) on page 189 of PRML. Change Eq. (4.20) into $y = \mathbf{w}^\mathsf{T} \boldsymbol{\phi}(\mathbf{x})$ and derive a 'kernelized' version of Fisher' linear discriminant for two classes. Note that the kernelized version involves only kernel evaluations. The implicit function $\boldsymbol{\phi}(\mathbf{x})$ should not appear in the final result.

   Solution:

   Problem 4 of assignment 4.

5. **(40 points)** ν-SV regression

   Consider the following primal optimization problem in which C is a regularization constant and $\nu \geq 0$:

$$\begin{aligned}
\text{minimize} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C \cdot \left(\nu\epsilon + \frac{1}{N} \sum_{n=1}^{N} (\xi_n + \widehat{\xi}_n) \right) \\
\text{subject to} \quad & t_n \leq y(\mathbf{x}_n) + \epsilon + \xi_n \,, \\
& t_n \geq y(\mathbf{x}_n) - \epsilon - \widehat{\xi}_n \,, \\
& \xi_n \geq 0,\ \widehat{\xi}_n \geq 0,\ \text{and } \epsilon \geq 0 \,, \\
& n = 1, \ldots, N \,.
\end{aligned}$$

(1) Introduce multipliers a_n, $\widehat{a}_n$, μ_n, $\widehat{\mu}_n$, and β for the respective constraints, and write down the Lagrangian function.

(2) Set the derivatives with respect to the primal variables equal to zero, and write down the corresponding four equations.

(3) Derive the dual problem.

(4) Write down the corresponding KKT conditions, and give a brief analysis on the results.

Solution:

(1)

$$\begin{aligned}L = \frac{1}{2}\|\mathbf{w}\|^2 + C\nu\epsilon + \frac{C}{N}\sum_{n=1}^{N}(\xi_n + \widehat{\xi}_n) - \beta\epsilon - \sum_{n=1}^{N}(\mu_n\xi_n + \widehat{\mu}_n\widehat{\xi}_n) \\ - \sum_{n=1}^{N} a_n(y(\mathbf{x}_n) + \epsilon + \xi_n - t_n) - \sum_{n=1}^{N}\widehat{a}_n(t_n - y(\mathbf{x}_n) + \epsilon + \widehat{\xi}_n)\,.\end{aligned}$$

(2)

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{w}} &\Rightarrow \mathbf{w} = \sum_{n=1}^{N}(a_n - \widehat{a}_n)\boldsymbol{\phi}(\mathbf{x}_n)\,,\\
\frac{\partial L}{\partial b} &\Rightarrow \sum_{n=1}^{N}(a_n - \widehat{a}_n) = 0\,,\\
\frac{\partial L}{\partial \epsilon} &\Rightarrow C\,\nu - \sum_{n=1}^{N}(a_n + \widehat{a}_n) - \beta = 0\,,\\
\frac{\partial L}{\partial \xi_n} &\Rightarrow a_n + \mu_n = \frac{C}{N}\,,\ n = 1, \dots, N,\\
\frac{\partial L}{\partial \widehat{\xi}_n} &\Rightarrow \widehat{a}_n + \widehat{\mu}_n = \frac{C}{N}\,,\ n = 1, \dots, N.
\end{aligned}$$

(3)

$$
\begin{aligned}
\text{maximize} \quad & \sum_{n=1}^{N}(a_n - \widehat{a}_n)t_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(a_n - \widehat{a}_n)(a_m - \widehat{a}_m)k(\mathbf{x}_n, \mathbf{x}_m) \\
\text{subject to} \quad & \sum_{n=1}^{N}(a_n - \widehat{a}_n) = 0\,, \\
& \sum_{n=1}^{N}(a_n + \widehat{a}_n) \le C\,\nu\,, \\
& 0 \le a_n \le \frac{C}{N}\,,\; 0 \le \widehat{a}_n \le \frac{C}{N}\,,\; n = 1, \ldots, N\,.
\end{aligned}
$$

(4)

$$
\begin{aligned}
& a_n(\epsilon + \xi_n + y(\mathbf{x}_n) - t_n) = 0\,, \\
& \widehat{a}_n(\epsilon + \widehat{\xi}_n - y(\mathbf{x}_n) + t_n) = 0\,, \\
\mu\xi_n = 0 \quad \Rightarrow \quad & (\frac{C}{N} - a_n)\xi_n = 0\,, \\
\widehat{\mu}\widehat{\xi}_n = 0 \quad \Rightarrow \quad & (\frac{C}{N} - \widehat{a}_n)\widehat{\xi}_n = 0\,, \\
\beta\,\epsilon = 0 \quad \Rightarrow \quad & \left(C\,\nu - \sum_{n=1}^{N}(a_n + \widehat{a}_n)\right)\,\epsilon = 0\,.
\end{aligned}
$$

Observations:

i) If $\xi_n > 0$, then $a_n = C/N$. If $\widehat{\xi}_n > 0$, then $\widehat{a}_n = C/N$.

ii) For every data point $\mathbf{x}_n$, either a_n or $\widehat{a}_n$ must be zero. If $\nu > 1$, then ϵ must be zero.

iii) $\frac{\text{(\# of errors)}}{N} \le \nu$.

iv) $\frac{\text{(\# of SVs)}}{N} \ge \nu$.