CSCI 5525: Machine Learning

Final Exam

Jingxiang Li

December 17, 2015

Problem 1.a

Knowing that

$$p(C_i|x) = \frac{p(x|C_i)p(C_i)}{p(x)}, \quad i = 1, 2$$

we have

$$\frac{p(C_1|x)}{p(C_2|x)} = \frac{p(x|C_1)}{p(x|C2)} \cdot \frac{P(C_1)}{P(C_2)}$$

i.e.

$$\log\left(\frac{p(C_1|x)}{p(C_2|x)}\right) = \log P(x|C_1) + \log P(C_1) - \log P(x|C_2) - \log P(C_2)$$

Given

$$P(x|C_i) = \exp(\eta_i^T x)g(\eta_i)h(x), \quad i = 1, 2$$

The above formula becomes

$$\begin{aligned}\log\left(\frac{p(C_1|x)}{p(C_2|x)}\right) &= \eta_1^T x + \log g(\eta_1) + \log h(x) + \log P(C_1) - (\eta_2^T x + \log g(\eta_2) + \log h(x) + \log P(C_2)) \\ &= (\eta_1 - \eta_2)^T x + \log g(\eta_1) - \log g(\eta_2) + \log P(C_1) - \log P(C_2)\end{aligned}$$

Let $\boldsymbol{w} = \eta_1 - \eta_2$, $w_0 = \log g(\eta_1) - \log g(\eta_2) + \log P(C_1) - \log P(C_2)$

We have

$$\log\left(\frac{p(C_1|x)}{p(C_2|x)}\right) = \boldsymbol{w}^T x + w_0$$

Problem 1.b

Note that $E(w)$ can be separated into two parts. Let $E(w) = \ell(w) + \lambda||w||_1$, then it's sufficient to show that both $\ell(w)$ and $||w||_1$ are convex functions of w.

First, I will prove the convexity of $\ell(w)$ by checking its second derivative.

$$\nabla\ell(w) = \sum_{i=1}^{n} -y_i x_i + \frac{\exp(w^T x_i)}{1 + \exp(w^T x_i)} x_i$$

Let $\pi_i = \frac{\exp(w^T x_i)}{1+\exp(w^T x_i)}$, we have

$$\nabla\ell(w) = \sum_{i=1}^{n} x_i(\pi_i - y_i)$$

Then

$$\nabla^2\ell(w) = \sum_{i=1}^{n} x_i \frac{\exp(w^T x_i)}{[1 + \exp(w^T x_i)]^2} x_i^T = \sum_{i=1}^{n} x_i \pi_i (1 - \pi_i) x_i^T$$

Note that $0 < \pi_i < 1$. Therefore we can define $p_i = \sqrt{\pi_i(1 - \pi_i)}$, and

$$\nabla^2\ell(w) = \sum_{i=1}^{n} p_i x_i (p_i x_i)^T$$

Then $\forall \alpha \in \mathbb{R}^d$, $\alpha \neq 0$ we have

$$\alpha^T \nabla^2\ell(w)\alpha = \sum_{i=1}^{n} \alpha^T p_i x_i (p_i x_i)^T \alpha = \sum_{i=1}^{n} (\alpha^T p_i x_i)^2 \geq 0$$

which suggests that $\nabla^2 \ell(w)$ is positive semidefinite, i.e. $\ell(w)$ is convex.

Next I will prove the convexity of $||w||_1$, where $w \in \mathbb{R}^d$

First I will show $f(x) = |x|$ is convex. This can be done by the definition of convexity. Let $\forall x_1, x_2 \in \mathbb{R}^1$, $\forall \lambda \in [0, 1]$, here we consider two cases:

1. If x_1 and x_2 have the same sign, then obviously $f(\lambda x_1 + (1 - \lambda)x_2) = \lambda f(x_1) + (1 - \lambda)f(x_2)$.
2. If x_1 and x_2 have different sign, without loss generality, let's assume $x_1 \leq 0$ and $x_2 > 0$. If $|\lambda x_1 + (1 - \lambda)x_2| > 0$, then $|\lambda x_1 + (1 - \lambda)x_2| - \lambda|x_1| - (1 - \lambda)|x_2| = 2\lambda x_1 \leq 0$; otherwise $|\lambda x_1 + (1 - \lambda)x_2| < 0$, then $|\lambda x_1 + (1 - \lambda)x_2| - \lambda|x_1| - (1 - \lambda)|x_2| = -2(1 - \lambda)x_2 < 0$.

Combining the two cases, we have $\forall x_1, x_2 \in \mathbb{R}^1$, $\forall \lambda \in [0, 1]$, $f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2)$, suggesting that $f(x) = |x|$ is convex.

Then for $||w||_1 = \sum_{i=1}^d |w_i|$, this is simply the sum of convex functions. Hence $||w||_1$ is still convex.

Since both $\ell(w)$ and $||w||_1$ are convex functions of w, $E(w) = \ell(w) + \lambda||w||_1$, $\lambda > 0$ is convex.

Problem 2.a

$$\Lambda(x, u, v) = x^T P x + q^T x + u^T (Ax - a) + v^T (Bx - b)$$

where $u \in \mathbb{R}^{k_1}$, $v \in \mathbb{R}^{k_2}$, $v \geq 0$

Problem 2.b

Since P is positive definite, $\Lambda(x, u, v)$ has global minimum. We can find it by calculating the first derivative of Λ and set it to 0.

$$\frac{\partial \Lambda(x, u, v)}{\partial x} = Px + q + A^T u + B^T v := 0$$

i.e.

$$x^* = -P^{-1}(q + A^T u + B^T v)$$

Problem 2.c

To utilize the ADMM, we first introduce an extra parameter z, and let $Bx = z$. The original objective function can be written as

$$\min_{x,z} x^T P x + q^T x \quad \text{s.t.} \quad Ax = a,\ Bx = z,\ z \leq b$$

Let $g(z)$ be the indicator function of $z \leq b$. Then the augmented Lagrangian becomes

$$L_\rho = x^T P x + q^T x + g(z) + u^T (Ax - a) + v^T (Bx - z) + \frac{\rho_u}{2} ||Ax - a||^2 + \frac{\rho_v}{2} ||Bx - z||^2$$

where $u \in \mathbb{R}^{k_1}$, $v \in \mathbb{R}^{k_2}$

Here we need to derive the update function for x and z.

$$\frac{\partial L_\rho}{\partial x} = Px + q + A^T u + B^T v + \rho_u A^T (Ax - a) + \rho_v B^T (Bx - z) := 0$$

$$(P + \rho_u A^T A + \rho_v B^T B)x = \rho_u A^T a + \rho_v B^T z - q - A^T u - B^T v$$

$$x^* = (P + \rho_u A^T A + \rho_v B^T B)^{-1}(\rho_u A^T a + \rho_v B^T z - q - A^T u - B^T v)$$

Suppose $z \leq b$, then

$$\frac{\partial L_\rho}{\partial z} = -v + \rho_v (z - Bx) := 0$$

$$z^* = Bx + \frac{v}{\rho_v}$$

we set $z^* = b$ if $z > b$, i.e.

$$z^* = \min\left\{Bx + \frac{v}{\rho_v}, b\right\}$$

Then the update functions for ADMM are as follows:

$$
\begin{aligned}
x_{k+1} &= (P + \rho_u A^T A + \rho_v B^T B)^{-1}(\rho_u A^T a + \rho_v B^T z_k - q - A^T u_k - B^T v_k) \\
z_{k+1} &= \min \left\{ B x_{k+1} + \frac{v_k}{\rho_v}, b \right\} \\
u_{k+1} &= u_k + \rho_u (A x_{k+1} - a) \\
v_{k+1} &= v_k + \rho_v (B x_{k+1} - z_{k+1})
\end{aligned}
$$

Problem 3.a

$$
\begin{aligned}
p(z_k|x_n) &= \frac{p(x_n|z_k)p(z_k)}{p(x_n)} \\
&= \frac{p(x_n|z_k)p(z_k)}{\sum_{j=1}^{K} p(x_n|z_j)p(z_j)} \\
&= \frac{(2\pi)^{-\frac{d}{2}}|\Sigma|^{-\frac{1}{2}} \exp(-\frac{1}{2}(x_n-\mu_k)^T\Sigma^{-1}(x_n-\mu_k))\pi_k}{\sum_{j=1}^{K}(2\pi)^{-\frac{d}{2}}|\Sigma|^{-\frac{1}{2}} \exp(-\frac{1}{2}(x_n-\mu_j)^T\Sigma^{-1}(x_n-\mu_j))\pi_j} \\
&= \frac{\exp(-\frac{1}{2}(x_n-\mu_k)^T\Sigma^{-1}(x_n-\mu_k))\pi_k}{\sum_{j=1}^{K}\exp(-\frac{1}{2}(x_n-\mu_j)^T\Sigma^{-1}(x_n-\mu_j))\pi_j}
\end{aligned}
$$

Problem 3.b

Let $p_{i,k} = p(z_k|x_i)$, then we know the likelihood for X is

$$L = \prod_{i=1}^{N}\sum_{k=1}^{K} p(x_i|z_k)\pi_k$$

and the log-likelihood

$$l = \sum_{i=1}^{N}\log\sum_{k=1}^{K} p(x_i|z_k)\pi_k$$

Next I will take derivative w.r.t. μ_k and set it to 0

$$
\begin{aligned}
0 &= \frac{\partial l}{\partial \mu_k} \\
0 &= \sum_{i=1}^{N}\frac{\pi_k\frac{\partial p(x_i|z_k)}{\partial \mu_k}}{\sum_{j=1}^{K} p(x_i|z_j)\pi_j} \\
0 &= \sum_{i=1}^{N}\frac{p(x_i|z_k)\pi_k}{\sum_{j=1}^{K} p(x_i|z_j)\pi_j}\frac{\partial\left\{-\frac{1}{2}(x_i-\mu_k)^T\Sigma^{-1}(x_i-\mu_k)\right\}}{\partial\mu_k} \\
0 &= \sum_{i=1}^{N} p_{i,k}\Sigma^{-1}(x_i-\mu_k) \\
0 &= \Sigma^{-1}\sum_{i=1}^{N} p_{i,k}(x_i-\mu_k)
\end{aligned}
$$

Note that Σ^{-1} is invertible, suggesting that $\dim(\mathrm{NULL}(\Sigma^{-1})) = 0$, and

$$\sum_{i=1}^{N} p_{i,k}(x_i-\mu_k) = 0$$

Let $N_k = \sum_{i=1}^{N} p_{i,k}$, we have

$$\hat{\mu}_k = \frac{1}{N_k}\sum_{i=1}^{N} p_{i,k}x_i = \frac{1}{N_k}\sum_{i=1}^{N} p(z_k|x_i)x_i$$

Next I will derive the update function for Σ. Let $\Omega = \Sigma^{-1}$ be the precision matrix of Gaussian distributions. I will take derivative w.r.t. Ω, set it to 0 and then derive the close form of Σ.

$$\begin{aligned}
0 &= \frac{\partial l}{\partial \Omega} \\
0 &= \sum_{i=1}^{N} \frac{\sum_{j=1}^{K} \pi_j \frac{\partial p(x_i|z_j)}{\partial \Omega}}{\sum_{j=1}^{K} p(x_i|z_j)\pi_j} \\
0 &= \sum_{i=1}^{N} \frac{\sum_{j=1}^{K} p(x_i|z_j)\pi_j \frac{1}{p(x_i|z_j)} \frac{\partial p(x_i|z_j)}{\partial \Omega}}{\sum_{j=1}^{K} p(x_i|z_j)\pi_j} \\
0 &= \sum_{i=1}^{N} \frac{\sum_{j=1}^{K} p(x_i|z_j)\pi_j \frac{\partial \log p(x_i|z_j)}{\partial \Omega}}{\sum_{j=1}^{K} p(x_i|z_j)\pi_j} \\
0 &= \sum_{i=1}^{N} \sum_{j=1}^{K} p_{i,j} \frac{\partial \log p(x_i|z_j)}{\partial \Omega} \\
0 &= \sum_{i=1}^{N} \sum_{j=1}^{K} p_{i,j} \frac{\partial \left\{ \frac{1}{2} \log |\Omega| - \frac{1}{2}(x_i - \mu_j)^T \Omega (x_i - \mu_j) \right\}}{\partial \Omega} \\
0 &= \sum_{i=1}^{N} \sum_{j=1}^{K} p_{i,j} \Omega^{-1} - (x_i - \mu_j)(x_i - \mu_j)^T \\
\sum_{i=1}^{N} \sum_{j=1}^{K} p_{i,j} \Omega^{-1} &= \sum_{i=1}^{N} \sum_{j=1}^{K} p_{i,j} (x_i - \mu_j)(x_i - \mu_j)^T \\
\Omega^{-1} &= \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} p_{i,j} (x_i - \mu_j)(x_i - \mu_j)^T
\end{aligned}$$

Note that $\Omega = \Sigma^{-1}$, then we have

$$\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} p_{i,j} (x_i - \mu_j)(x_i - \mu_j)^T = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} p(z_j|x_i)(x_i - \mu_j)(x_i - \mu_j)^T$$

Next I will derive the update function for π_k. Since $\sum_{j=1}^{K} \pi_j = 1$, the optimization problem becomes

$$\max_{\pi_k} \sum_{i=1}^{N} \log \sum_{j=1}^{K} p(x_i|z_j)\pi_j \quad \text{s.t.} \quad \sum_{j=1}^{K} \pi_j = 1$$

The Lagrangian of this problem is

$$\Lambda(\pi_k, \lambda) = \sum_{i=1}^{N} \log \sum_{j=1}^{K} p(x_i|z_j)\pi_j + \lambda \left(1 - \sum_{j=1}^{K} \pi_j \right)$$

Then I will take the derivative w.r.t. π_k and set it to 0. Let $\sum_{i=1}^{N} p_{i,k} = N_k$

$$
\begin{aligned}
0 &= \frac{\partial \Lambda(\pi_k, \lambda)}{\partial \pi_k} \\
0 &= \sum_{i=1}^{N} \frac{p(x_i|z_k)}{\sum_{j=1}^{K} p(x_i|z_j)\pi_j} - \lambda \\
\lambda &= \frac{1}{\pi_k} \sum_{i=1}^{N} \frac{p(x_i|z_k)\pi_k}{\sum_{j=1}^{K} p(x_i|z_j)\pi_j} \\
\pi_k &= \frac{1}{\lambda} \sum_{i=1}^{N} p_{i,k} \\
\pi_k &= \frac{1}{\lambda} N_k
\end{aligned}
$$

Note that $\sum_{j=1}^{K} \pi_j = 1$, there must be $\lambda = N = \sum_{j=1}^{K} N_j$. By the strong duality of Lagrangian, we have

$$
\hat{\pi}_k = \frac{N_k}{N}
$$

In summary, the update functions for μ_k, π_k and Σ are given by

$$
\begin{aligned}
\hat{\mu}_k &= \frac{1}{N_k} \sum_{i=1}^{N} p(z_k|x_i)x_i \\
\hat{\Sigma} &= \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} p(z_j|x_i)(x_i - \mu_j)(x_i - \mu_j)^T \\
\hat{\pi}_k &= \frac{N_k}{N}
\end{aligned}
$$

Problem 3.c

No. If the covariance matrix for all Gaussian distributions are different, the update function for Σ_k would be

$$
\hat{\Sigma}_k = \frac{1}{N_k} \sum_{i=1}^{N} p(z_k|x_i)(x_i - \mu_k)(x_i - \mu_k)^T
$$

and

$$
\hat{\Sigma} = \frac{1}{N} \sum_{j=1}^{K} N_j \hat{\Sigma}_j
$$

suggesting that the computation for deriving $\hat{\Sigma}$ and all $\hat{\Sigma}_k$ are exactly the same. the computation of Σ is not simplified.

Problem 4.a

Given that $f(x) \sim \mathrm{GP}(m(x), k(x, x'))$, $y|f(x) \sim N(f(x), \sigma^2)$, by the conditional probability formula for joint Gaussian distribution, we have

$$f(x^*)|x^*, X, y \sim N(\mu_{f(x^*)}, \Sigma_{f(x^*)})$$

where

$$\mu_{f(x^*)} = m(x^*) + k(x^*, X)[k(X, X) + \sigma^2]^{-1}(y - m(X))$$

and

$$\Sigma_{f(x^*)} = k(x^*, x^*) - k(x^*, X)[k(X, X) + \sigma^2]^{-1}k(X, x^*)$$

Then we have

$$y^*|x^*, X, y \sim N(\mu_{y^*}, \Sigma_{y^*})$$

where

$$\mu_{y^*} = m(x^*) + k(x^*, X)[k(X, X) + \sigma^2]^{-1}(y - m(X))$$

and

$$\Sigma_{y^*} = \sigma^2 + k(x^*, x^*) - k(x^*, X)[k(X, X) + \sigma^2]^{-1}k(X, x^*)$$

Then it's easy to see that

$$\mu_{y^*} = m(x^*) + k(x^*, X)[k(X, X) + \sigma^2]^{-1}(y - m(X)) = \beta_0 + \sum_{i=1}^{n} \beta_i y_i = \alpha_0 + \sum_{i=1}^{n} \alpha_i k(x^*, X_i)$$

where

$$\beta_0 = m(x^*) - k(x^*, X)[k(X, X) + \sigma^2]^{-1}m(X) \qquad \beta_i = \left[k(x^*, X)[k(X, X) + \sigma^2]^{-1}\right]_i$$

$$\alpha_0 = m(x^*) \qquad \alpha_i = \left[[k(X, X) + \sigma^2]^{-1}(y - m(X))\right]_i$$

Problem 4.b

It depends on the choice of the mean function m, the kernel function k and the training set X. Suppose $n = 1$, $m(x) = y_1$, then

$$\begin{aligned}
\mu_{y^*} &= m(x^*) + k(x^*, X)[k(X, X) + \sigma^2]^{-1}(y - m(X)) \\
&= y_1 + k(x^*, x_1)[k(x_1, x_1) + \sigma^2]^{-1}(y_1 - y_1) \\
&= y_1
\end{aligned}$$

In this case the mean of the predictive distribution exactly overlaps with y_i.

However, in general case this would not happen for the following reasons:

1. the mean of the predictive distribution will be dragged to the mean function $m(x)$
2. the mean will be influenced by other points x', if $k(x^*, x') \neq 0$

In summary, the predictive distribution at x^* will exactly overlap with y_i in some special cases, but in general cases this would not happen.

Problem 4.c

Yes, we can obtain a closed form expression for the m-dimensional predictive joint distribution. By the definition of Gaussian process, y^* and y will have joint Gaussian distribution, the mean and the covariance matrix are clearly defined by the mean function $m(x)$ and kernel function $k(x, x')$. Then we can use the conditional probability formulas

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2) \qquad \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

to derive the predictive distribution.