ENGG 5202: Midterm Solution

March 2014

1 Problem 1

1. False.
2. False.
3. True, when the number of training samples is very large, $P(\theta|D)$ approaches a delta function $\delta(\theta-\hat{\theta})$, and Bayesian estimation becomes equivalent to the Maximum Likelihood estimation. In Bayesian estimation, $p(x|D) = \int p(x|\theta)p(\theta|D)d\theta = \int p(x|\theta)\delta(\theta-\hat{\theta})d\theta = p(x|\hat{\theta})$.
4. False
5. False. Since there are enough training samples, high-order polynomial coefficients will approach zeros. Thus, 10th degree polynomial will not lead to overfitting problem.
6. True. Null space LDA projects training sample to a subspace where within-class variance is zero, which means samples of each class will be projected to a single point, $w^t\bar{x}_1$ and $w^t\bar{x}_2$ in this case. And if $w^t\bar{x}_1 \neq w^t\bar{x}_2$, the two projected centers are different, so we can completely separate two classes, and achieve zero classification error.
7. False.
8. False. Given x_t, x_{t-1} and x_{t+1} are not blocked since there are path through the hidden variables, thus, x_{t-1} and x_{t+1} are not independent given x_t.
9. True. Since the nearest neighbor error rate is bounded by Bayes error rate, i.e. $P^* \leq P \leq P^*(2-\frac{c}{c-1}P^*)$, when $P^* = 0$, we have $P_n = P = 0$ given infinite training samples.
10. True.

2 Problem 2

(1) The discriminative function is

$$g_i(x) = \ln p(\mathbf{x}|w_i) + \ln P(w_i) \tag{1}$$

$$p(\mathbf{x}|w_i) \sim \mathcal{N}(\mu_i, \mathbf{\Sigma}_i),\ \Sigma_1 = \Sigma_2 = \Sigma \tag{2}$$

$$\Rightarrow g_i(x) = -\frac{1}{2}(\mathbf{x}-\mu_i)^t\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{\Sigma}| + \ln P(w_i)$$

Ignore terms unrelated to class i, we have

$$g_i(x) = \mathbf{x}^t \mathbf{\Sigma}^{-1} \mu_i - \frac{1}{2} \mu_i^t \mathbf{\Sigma}^{-1} \mu_i + \ln P(w_i) \tag{3}$$

Decision boundary is the hyperplane defined by $g_1(\mathbf{x}) = g_2(\mathbf{x})$,

$$\mathbf{x}^t \mathbf{\Sigma}^{-1} \mu_1 - \frac{1}{2} \mu_1^t \mathbf{\Sigma}^{-1} \mu_1 + \ln P(w_1) = \mathbf{x}^t \mathbf{\Sigma}^{-1} \mu_2 - \frac{1}{2} \mu_2^t \mathbf{\Sigma}^{-1} \mu_2 + \ln P(w_2) \tag{4}$$

$$\Rightarrow (\mathbf{\Sigma}^{-1}(\mu_1 - \mu_2))^t \mathbf{x} - \Big(\frac{1}{2} (\mu_1 - \mu_2)^t \mathbf{\Sigma}^{-1} (\mu_1 + \mu_2) - \ln \frac{P(w_1)}{P(w_2)} \Big) = 0 \tag{5}$$

$$\Leftrightarrow \mathbf{w}^t (\mathbf{x} - \mathbf{x}_0) = 0 \tag{6}$$

where

$$\mathbf{w} = \mathbf{\Sigma}^{-1}(\mu_1 - \mu_2), \quad \mathbf{x}_0 = \frac{1}{2}(\mu_1 + \mu_2) - \frac{\ln[P(w_1)/P(w_2)](\mu_1 - \mu_2)}{(\mu_1 - \mu_2)^t \mathbf{\Sigma}^{-1} (\mu_1 - \mu_2)} \tag{7}$$

(2) To make the decision boundary passing through middle of two class means,

i.e. $\mathbf{x}_0 = \frac{1}{2}(\mu_1 + \mu_2)$.

$$\Rightarrow \frac{\ln[P(w_1)/P(w_2)](\mu_1 - \mu_2)}{(\mu_1 - \mu_2)^t \mathbf{\Sigma}^{-1} (\mu_1 - \mu_2)} = 0 \tag{8}$$

Because Σ is positive definite, and $\mu_1 \neq \mu_2$, we know that $(\mu_1 - \mu_2)^t \mathbf{\Sigma}^{-1} (\mu_1 - \mu_2) > 0$, thus, the condition is $P(w_1) = P(w_2)$.

(3) To make the $\mathbf{w}$ in the same direction as $\mu_1 - \mu_2$, i.e.

$$\mathbf{w} = \mathbf{\Sigma}^{-1}(\mu_1 - \mu_2) = \lambda(\mu_1 - \mu_2) \tag{9}$$

where $\lambda > 0$, this indicates that $\mu_1 - \mu_2$ is one of the eigenvector of Σ or Σ^{-1} with non-zero eigenvalue.

(4) Not possible. If $\mathbf{w}$ is orthogonal to $\mu_1 - \mu_2$, then

$$(\mu_1 - \mu_2)^t \mathbf{w} = (\mu_1 - \mu_2)^t \mathbf{\Sigma}^{-1} (\mu_1 - \mu_2) = 0 \tag{10}$$

But Σ is positive definite and so is Σ^{-1}, thus we have $(\mu_1 - \mu_2)^t \mathbf{\Sigma}^{-1} (\mu_1 - \mu_2) > 0$ when $\mu_1 \neq \mu_2$, which is a contradiction.

3 Problem 3

Likelihood of each class is

$$P(D_i|\theta_i) = \prod_{k=1}^{2} \frac{2}{\theta_i} (1 - \frac{x_k}{\theta_i}), \quad \theta_i \geq \max_k x_k \tag{11}$$

For class 1:

$$P(D_1|\theta_1) = \frac{4}{\theta_1^2}(1-\frac{2}{\theta_1})(1-\frac{5}{\theta_1}), \quad \theta_1 \geq 5 \tag{12}$$

Let $t_1 = \frac{1}{\theta_1}$, we have

$$P(D_1|t_1) = 4t_1^2(1-2t_1)(1-5t_1), \quad 0 < t_1 \leq \frac{1}{5} \tag{13}$$

$$\frac{\partial P(D_1|t_1)}{\partial t_1} = 160t_1^3 - 84t_1^2 + 8t_1 = 0, \quad 0 < t_1 \leq \frac{1}{5} \tag{14}$$

It is easy to see $t_1 = 0,\ \frac{1}{8}$, or $\frac{2}{5}$. With $0 < t_1 \leq \frac{1}{5}$, we get $t_1 = \frac{1}{8}$, then $\theta_1 = 8$. Similarly for class 2:

$$P(D_2|\theta_2) = \frac{4}{\theta_2^2}(1-\frac{3}{\theta_2})(1-\frac{9}{\theta_2}), \quad \theta_2 \geq 9 \tag{15}$$

Let $t_2 = \frac{1}{\theta_2}$, we have

$$P(D_2|t_2) = 4t_2^2(1-3t_2)(1-9t_2), \quad 0 < t_2 \leq \frac{1}{9} \tag{16}$$

$$\frac{\partial P(D_2|t_2)}{\partial t_2} = 4(108t_2^3 - 36t_2^2 + 2t_2) = 0, \quad 0 < t_2 \leq \frac{1}{9} \tag{17}$$

It is easy to see $t_2 = 0$ or $\frac{3\pm\sqrt{3}}{18}$. With $0 < t_1 \leq \frac{1}{9}$, we get $t_2 = \frac{3-\sqrt{3}}{18}$, then $\theta_2 = \frac{18}{3-\sqrt{3}} = 9 + 3\sqrt{3}$.

4 Problem 4

For each EM iteration, estimation of a_{12} is updated in the M step as follow

$$\hat{a}_{12} = \frac{\sum_{t=2}^{T} \xi_{t-1}(w_1, w_2)}{\sum_{t=2}^{T}\sum_{j'=1}^{c} \xi_{t-1}(w_1, w_{j'})} \tag{18}$$

where $\xi_t(w_1, w_2)$ depends on the initialization of a_{12}

$$\xi_t(w_1, w_2) = \frac{\alpha_t(w_1)a_{12}P(x_{t+1}|z_{t+1} = w_2)\beta_{t+1}(w_2)}{\sum_{j'} \alpha_t(w_{j'})\beta_t(w_{j'})} \tag{19}$$

If a_{12} is initialized to be zero, then $\xi_t(w_1, w_2)$ will all be zeros for $t = 2, \ldots, T$. Therefore, a_{12} will remain zero in all subsequent EM updates. [Remark: refer to Tutorial 5 for detailed derivation]