UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
Department of Electrical and Computer Engineering

ECE 544NA Pattern Recognition
Fall 2014

EXAM 3

Monday, December 15, 2014

- This is a CLOSED BOOK exam. You may use two pages, both sides, of notes.
- There are a total of 100 points in the exam. Plan your work accordingly.
- You must SHOW YOUR WORK to get full credit.

Problem	Score
1	
2	
3	
4	
Total	

Name: ____

Problem 1 (25 points)

In this problem, the observation $x \in [0,1]$ is a real number drawn from a uniform distribution,

$$p_x(x) = \begin{cases} 1 & 0 \le x \le 1, \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The true label of each datum is $y = [x > \theta]$, where $[\cdot]$ is the unit indicator function, and θ is an unknown threshold parameter. Suppose that the prior distribution for θ is also uniform:

$$p_\theta(\theta) = \begin{cases} 1 & 0 \le \theta \le 1, \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The hypothesis space is the set of all threshold functions,

$$\mathcal{H} = \left\{ h(x) = \left[x > \hat{\theta} \right] : \quad \hat{\theta} \in [0,1] \right\}$$

The feasible set after training on a set of n labeled data is the set of all hypotheses that do not contradict any of the training data:

$$\mathcal{H}_n = \{ h : \quad h \in \mathcal{H}, \quad h(x_i) = y_i \; \forall \; 1 \le i \le n \}$$

The worst-case risk, after n training data, is

$$R_n = \max_{h \in \mathcal{H}_n} \Pr\left\{ y \ne h(x) \right\}$$

(a) Assume that x_i are drawn at random according to Eq. 1. Notice that, in this case, R_n is a random variable. Define its cumulative distribution function to be

$$F_n(\epsilon) = \Pr\left\{ R_n \ge \epsilon \right\}$$

Find $F_n(\epsilon)$ as a function of ϵ. You may assume that $\epsilon \le \theta \le 1 - \epsilon$.

(b) Suppose that you are allowed to use the following active learning algorithm.

(i) Set the base to $b_1 = 0$, the step to $s_1 = 0.5$.

(ii) For $1 \leq i \leq n$:

i. Set $x_i = b_i + s_i$. Ask a teacher to label this token, giving the true value of y_i.

ii. If $y_i == 0$, set the base to $b_{i+1} = x_i$, else $b_{i+1} = b_i$.

iii. $s_{i+1} = s_i/2$.

R_n is still a random variable (because of Eq. 2), but now it has a much reduced range. Find $F_n(\epsilon)$.

Problem 2 (25 points)

K-means clustering finds a set of modes, $\theta = \{\mu_1, \ldots, \mu_K\}$, in order to minimize

$$\mathcal{E} = \sum_{i=1}^{n} \|x_i - \mu_{k_i}\|^2$$

where k_i is the cluster assignment of the i^{th} training datum. The K-means algorithm progressively reduces $\mathcal{E}$ by iteratively alternating between Eq. 3 and Eq. 4:

$$k_i = \arg\min \|x_i - \mu_k\| \quad (3)$$

$$\mu_k = \frac{1}{n_k} \sum_{i:k_i=k} x_i \quad (4)$$

where n_k is the number of data for which $k_i = k$.

(a) Prove that Eq. 4 minimizes $\mathcal{E}$ for fixed values of k_i.

(b) Suppose that you have a semi-supervised learning problem in which there are n labeled data (x_1 through x_n), and u unlabeled data (x_{n+1} through x_{n+u}). Suppose that you decide to minimize the joint criterion

$$\mathcal{F} = \sum_{i=1}^{n+u} \|x_i - \mu_{k_i}\|^2 + \lambda \sum_{i=1}^{n} \frac{[y_i \neq y(k_i)]}{n_k} \tag{5}$$

where $[\cdot]$ is the unit indicator function, $\lambda > 0$ is some real-valued regularizing parameter, and $y(k)$ is the majority class label of cluster k defined as

$$y(k) = \text{argmax}_y \sum_{i:k_i=k} [y_i = y]$$

It is possible to create a version of the K-means algorithm that progressively minimizes Eq. 5. Indeed, Eq. 4 reduces $\mathcal{F}$ in exactly the same way that it minimizes $\mathcal{E}$. Eq. 3, however, needs to be modified.

Suppose that each training datum has a previous cluster affiliation, $\hat{k}_i$. Your goal is to create a new cluster affiliation k_i that changes ($k_i \neq \hat{k}_i$) if and only if a change will reduce $\mathcal{F}$, thus

$$k_i = \arg\min \mathcal{F} \quad \text{s.t. } k_j = \hat{k}_j \text{ for all } j \neq i$$

Find the condition under which $k_i \neq \hat{k}_i$. Your condition will depend on the value of λ.

Problem 3 (25 points)

Suppose that you have a problem characterized by non-negative real observations, that is, $v \in \Re^+$. Consider a mixture exponential hypothesis:

$$p_v(v) = \begin{cases} \sum_{h=1}^{m} c_h \lambda_h e^{-\lambda_h v} & v \geq 0 \\ 0 & v < 0 \end{cases} \tag{6}$$

where $\lambda_h > 0$ is the rate of the h^{th} exponential, $c_h \geq 0$, and $1 = \sum_{h=1}^{m} c_h$.

Define the trainable parameters $\theta = \{c_1, \lambda_1, \ldots, c_m, \lambda_m\}$, and define the posterior probability

$$\gamma_i(h; \theta) = p_{h|v}(h|v_i, \theta)$$

(a) Write $\gamma_i(h; \theta)$ as an explicit function of v_i and of the trainable parameters.

(b) Define

$$Q(\theta, \hat{\theta}) = \sum_{i=1}^{n} \sum_{h} \gamma_i(h; \hat{\theta}) \ln p_{h,v}(h, v_i|\theta)$$

In terms of $\gamma_i(h; \hat{\theta})$ and v_i, find the value of λ_h that maximizes $Q(\theta, \hat{\theta})$.

Problem 4 (25 points)

Consider a neural network defined by input vectors $x_i = [x_{i1}, \ldots, x_{ip}]^T$, targets $t_i = [t_{i1}, \ldots, t_{ir}]^T$, and by the following transformations

$$\begin{aligned} a_{ik} &= \sum_{j=1}^{p} u_{kj} x_{ij} && (7) \\ y_{ik} &= f(a_{ik}) && (8) \\ b_{i\ell} &= \sum_{k=1}^{q} v_{\ell k} y_{ik} && (9) \\ z_{i\ell} &= g(b_{i\ell}) && (10) \\ \mathcal{E} &= \sum_{i=1}^{n} \sum_{\ell=1}^{r} t_{i\ell} \ln \left(\frac{t_{i\ell}}{z_{i\ell}} \right) && (11) \end{aligned}$$

(a) Define

$$\epsilon_{i\ell} = \frac{\partial \mathcal{E}}{\partial b_{i\ell}}$$

Express $\epsilon_{i\ell}$ as an explicit function of $t_{i\ell}$, $z_{i\ell}$, and the derivative function $g'(b_{i\ell})$. You may assume that $t_{i\ell} \geq 0$, $z_{i\ell} > 0$, and $0 \ln 0 \equiv 0$.

(b) Find

$$\frac{\partial \mathcal{E}}{\partial v_{\ell k}} \quad \text{and} \quad \frac{\partial \mathcal{E}}{\partial u_{kj}}$$

as explicit functions of $\epsilon_{i\ell}$, y_{ik}, x_{ij}, and $v_{\ell k}$. If you need to define any other intermediate variables, make certain that you define them clearly.