Name: Jiaying Shi Student ID: 24978491

CS189: Introduction to Machine Learning

Homework 2

Due: February 17, 2015 at 11:59pm

Problem 1. A target is made of 3 concentric circles of radii $1/\sqrt{3}$, 1 and $\sqrt{3}$ feet. Shots within the inner circle are given 4 points, shots within the next ring are given 3 points, and shots within the third ring are given 2 points. Shots outside the target are given 0 points.

Let X be the distance of the hit from the center (in feet), and let the p.d.f of X be

$$f(x) = \begin{cases} \frac{2}{\pi(1+x^2)} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

What is the expected value of the score of a single shot?

Solution:

Let Y be the score of a single shot,

$\Pr(Y=4) = \Pr(X \le \frac{1}{\sqrt{3}}) = \int_0^{\frac{1}{\sqrt{3}}} f(x)dx = \frac{2}{\pi}\arctan(x)\Big|_0^{\frac{1}{\sqrt{3}}} = \frac{2}{\pi}\cdot\frac{\pi}{6} = \frac{1}{3}$

$\Pr(Y=3) = \Pr(\frac{1}{\sqrt{3}} < X \le 1) = \int_{\frac{1}{\sqrt{3}}}^{1} f(x)dx = \frac{2}{\pi}\arctan(x)\Big|_{\frac{1}{\sqrt{3}}}^{1} = \frac{2}{\pi}(\frac{\pi}{4}-\frac{\pi}{6}) = \frac{1}{6}$

$\Pr(Y=2) = \Pr(1 < X \le \sqrt{3}) = \int_1^{\sqrt{3}} f(x)dx = \frac{2}{\pi}\arctan(x)\Big|_1^{\sqrt{3}} = \frac{2}{\pi}(\frac{\pi}{3}-\frac{\pi}{4}) = \frac{1}{6}$

$\Pr(Y=0) = 1 - \Pr(Y=4) - \Pr(Y=3) - \Pr(Y=2) = \frac{1}{3}$

Let $\mathcal{Y} = \{0, 2, 3, 4\}$

$$E(Y) = \sum_{y\in\mathcal{Y}} y\Pr(Y=y) = 0\cdot\frac{1}{3} + 2\cdot\frac{1}{6} + 3\cdot\frac{1}{6} + 4\cdot\frac{1}{3} = \frac{13}{6} \doteq 2.167.$$

The expected value of the score of a single shot is 2.167.

Problem 2. Assume that the random variable X has the exponential distribution

$$f(x|\theta) = \theta e^{-\theta x} \qquad x > 0, \theta > 0$$

where θ is the parameter of the distribution. Use the method of maximum likelihood to estimate θ if 5 observations of X are $x_1 = 0.9$, $x_2 = 1.7$, $x_3 = 0.4$, $x_4 = 0.3$, and $x_5 = 2.4$, generated i.i.d.

Solution:

The likelihood function can be written as .

$$L(\theta) = \theta^5 e^{-\theta \sum_{i=1}^{5} x_i} = \theta^5 e^{-5.7\theta}.$$

$$\frac{\partial L(\theta)}{\partial \theta} = (5\theta^4 - 5.7\theta^5)\, e^{-5.7\theta} \qquad ①$$

$$\frac{\partial^2 L(\theta)}{\partial \theta^2} = 20\theta^3 e^{-5.7\theta} + \theta^5 \cdot 5.7^2 e^{-5.7\theta} > 0 \quad ②.$$

By ②, we know $L(\theta)$ is concave and θ^* such that $\frac{\partial L(\theta)}{\partial \theta}\Big|_{\theta=\theta^*} = 0$ will maximize $L(\theta)$

$$\frac{\partial L(\theta)}{\partial \theta} = 0 \Rightarrow 5\theta^4 - 5.7\theta^5 = 0 \quad \text{since } \theta > 0 \quad \theta = \frac{50}{57} = 0.877$$

Problem 3. The polynomial kernel is defined to be

$$k(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^{\mathbf{T}}\mathbf{y} + \mathbf{c})^{\mathbf{d}}$$

where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{\mathbf{n}}$, and $c \geq 0$. When we take $d = 2$, this kernel is called the quadratic kernel.

(a) Find the feature mapping $\Phi(\mathbf{z})$ that corresponds to the quadratic kernel.

(b) How do we find the optimal value of d for a given dataset?

Solution:

(a) $k(x,y) = (x^Ty+c)^2 = (x^Ty)^2 + 2cx^Ty + c^2 = (x_1y_1+\cdots+x_ny_n)^2 + 2c(x_1y_1+\cdots x_ny_n)+c^2$

$$= \begin{pmatrix} c \\ \sqrt{2c}\,x_1 \\ \vdots \\ \sqrt{2c}\,x_n \\ x_1^2 \\ \vdots \\ x_n^2 \\ \sqrt{2}x_1x_2 \\ \sqrt{2}x_1x_3 \\ \vdots \\ \sqrt{2}x_{n-1}x_n \end{pmatrix} \cdot \begin{pmatrix} c \\ \sqrt{2c}\,y_1 \\ \vdots \\ \sqrt{2c}\,y_n \\ y_1^2 \\ \vdots \\ y_n^2 \\ \sqrt{2}y_1y_2 \\ \sqrt{2}y_1y_3 \\ \vdots \\ \sqrt{2}y_{n-1}y_n \end{pmatrix}$$

$\phi(x) \cdot \phi(y)$.

Thus $\phi(z) = (c, \sqrt{2c}\,z_1, \cdots \sqrt{2c}\,z_n, z_1^2 \cdots z_n^2, \sqrt{2}z_1z_2, \sqrt{2}z_1z_3, \cdots \underset{i\neq j}{\sqrt{2}z_iz_j} \cdots \sqrt{2}z_{n-1}z_n)^T$

(b). We can use cross-validation to see which value of d has the best performance and avoid over-fitting.
We can devide the data set into k folds and use k-1 of them to train and use the remaining one to do the test. Record accuracy rate for each test and then take the average. accuracy rate as the performance measure of some value of d. Try different values of d and take the one with the best performance as the optimal d value.

Def: Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix. We say that A is positive definite if $\forall x \in \mathbb{R}^n$, $x^\top Ax > 0$. Similarly, we say that A is positive semidefinite if $\forall x \in \mathbb{R}^n$, $x^\top Ax \geq 0$.

Problem 4. Let $x = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top \in \mathbb{R}^n$, and let $A \in \mathbb{R}^{n\times n}$ be the square matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

(a) Give an explicit formula for $x^\top Ax$. Write your answer as a sum involving the elements of A and x.

Solution:

$$x^\top Ax = \begin{bmatrix} x_1, \cdots, x_n \end{bmatrix} \begin{bmatrix} \sum_{j=1}^n a_{1j} x_j \\ \sum_{j=1}^n a_{2j} x_j \\ \vdots \\ \sum_{j=1}^n a_{nj} x_j \end{bmatrix} = \sum_{i=1}^n x_i \sum_{j=1}^n a_{ij} x_j = \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j .$$

(b) Show that if A is positive definite, then the entries on the diagonal of A are positive (that is, $a_{ii} > 0$ for all $1 \leq i \leq n$).

Solution:

Assume $\exists\ i \in \{1, \cdots n\}$ $a_{ii} \leq 0$.

Let $\bar{x} = [0, \cdots 0, 1, 0, \cdots 0]$ where $\bar{x}_i = 1$ $\bar{x}_j = 0$ for $\forall j \neq i$.

$\bar{x}^\top A \bar{x} = a_{ii} \cdot 1^2 = a_{ii} \leq 0$. which contradicts to that A is positive definite.

That means $\nexists\ i \in \{1, \cdots n\}$ $a_{ii} \leq 0$.

Thus $\forall\ i \in \{1, \cdots n\}$ $a_{ii} > 0$.

Problem 5. Let B be a positive semidefinite matrix. Show that $B + \gamma I$ is positive definite for any $\gamma > 0$.

Solution:

B is positive semidefinite $\Rightarrow \forall x \neq 0,\ x \in R^n \quad x^T B x \geq 0$.

$\forall x \neq 0,\ x \in R^n$.

$x^T(B+\lambda I)x = x^T B x + \gamma x^T I x = x^T B x + \gamma x^T x$

$\gamma > 0,\ x \neq 0 \Rightarrow \gamma x^T x > 0$.

$x^T B x \geq 0$.

Thus. $\forall x \neq 0\ x \in R^n$ we have $x^T(B+\lambda I)x = x^T B x + \gamma x^T x > 0$.

Then $B+\gamma I$ is positive definite for any $\gamma > 0$ by definition.

Problem 6. Suppose we have a classification problem with classes labeled $1, \ldots, c$ and an additional doubt category labeled as $c+1$. Let the loss function be the following:

$$\ell(f(x) = i, y = j) = \begin{cases} 0 & \text{if } i = j \quad i, j \in \{1, \ldots, c\} \\ \lambda_r & \text{if } i = c+1 \\ \lambda_s & \text{otherwise} \end{cases}$$

where λ_r is the loss incurred for choosing doubt and λ_s is the loss incurred for making a misclassification. Note that $\lambda_r \geq 0$ and $\lambda_s \geq 0$.

(a) Show that the minimum risk is obtained if we follow this policy: (1) choose class i if $P(\omega_i|x) \geq P(\omega_j|x)$ for all j and $P(\omega_i|x) \geq 1 - \lambda_r/\lambda_s$, and (2) choose doubt otherwise.

Solution:

Case1: $\exists i \neq c+1$ satisfies (1), according to the policy, we should choose i.

The expected value of risk is $E(\bar{l}) = \lambda_s \sum_{\substack{k=1 \\ k \neq i}}^{c+1} P_r(w_k|x) = \lambda_s (1 - P_r(w_i|x))$.

1) For $\forall n \neq i, c+1$, if choose n, the expected value of risk is $E(l) = \lambda_s (1 - P_r(w_n|x))$

$P_r(w_i|x) \geq P_r(w_n|x) \Rightarrow E(\bar{l}) \leq E(l)$

2) For $n = c+1$, if choose n. the expected value of risk is

$E(l) = \lambda_r \sum_{k=1}^{c+1} P_r(w_k|x) = \lambda_r$. $\quad P_r(w_i|x) \geq 1 - \frac{\lambda_r}{\lambda_s} \Rightarrow \frac{\lambda_r}{\lambda_s} \geq 1 - P_r(w_i|x)$

$E(\bar{l}) = \lambda_s (1 - P_r(w_i|x)) \leq \lambda_s \cdot \frac{\lambda_r}{\lambda_s} = \lambda_r = E(l)$

$E(l) \geq E(\bar{l})$ That means in this case choosing i will minimize the risk.

Case2: according to the policy we should choose $c+1$ which means

$\nexists i \in \{1, \ldots c\}$ $P_r(w_i|x) \geq P_r(w_j|x)$ $\forall j$ and $P_r(w_i|x) \geq 1 - \frac{\lambda_r}{\lambda_s}$ ⊛.

The expected value of risk of choosing $c+1$ is $E(\bar{l}) = \lambda_r$.

For $i \in \{1, \ldots c\}$ $\exists\, i_0$ $P_r(w_{i_0}|x) \geq P_r(w_j|x)$ $\forall j \in \{1 \ldots c\}$

The expected value of risk of choosing $j \in \{1 \ldots c\}$ is.

$E(l_j) = \lambda_s (1 - P_r(w_j|x)) > E(l_{i_0}) = \lambda_s (1 - P_r(w_{i_0}|x))$

By ⊛ we know $P_r(w_{i_0}|x) < 1 - \frac{\lambda_r}{\lambda_s}$ $\quad 1 - P_r(w_{i_0}|x) > \frac{\lambda_r}{\lambda_s} \Rightarrow E(l_{i_0}) > E(\bar{l})$

That means in this case choosing doubt will minimize the risk.

Thus from the analysis of the two cases, the minimum risk is obtained from the policy.

(b) What happens if $\lambda_r = 0$? What happens if $\lambda_r > \lambda_s$?
Solution:

If $\lambda_r = 0$. (1) will be satisfied only if $\exists i \; P(w_i|x) = 1$ in that case choose class i and in this case there is no need to classify. Otherwise. we will always choose doubt.

If $\lambda_r > \lambda_s$. the cost of choosing doubt is higher (1) will always be satisfied by some $i \neq c+1$ so doubt will never be chosen.

Problem 7. Let $p(x|\omega_i) \sim \mathcal{N}(\mu_i, \sigma^2)$ for a two-category one-dimensional classification problem with $P(\omega_1) = P(\omega_2) = 1/2$.

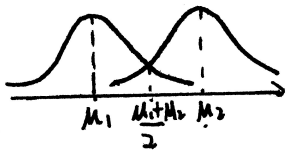
(a) Show that the minimum probability of error is

$$P_e = \frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-(1/2)u^2} du$$

where $a = |\mu_2 - \mu_1|/2\sigma$.

Solution:

Since $\sigma_1 = \sigma_2$ $P_r(\omega_1) = P_r(\omega_2)$, the boundry of the classification is $\frac{\mu_1+\mu_2}{2}$. without loss of generality, let's assume $\mu_1 < \mu_2$.



$$P_e = P_r(\omega_1)\, P_r\left(X > \tfrac{\mu_1+\mu_2}{2} \mid \omega_1\right) + P_r(\omega_2)\, P_r\left(X < \tfrac{\mu_1+\mu_2}{2} \mid \omega_2\right)$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma} \int_{\frac{\mu_1+\mu_2}{2}}^{+\infty} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}} dx \quad \text{(by symmetry)}.$$

Let $u = \frac{x-\mu_1}{\sigma}$ $\quad du = \frac{1}{\sigma} dx$. The range of the integration becomes $[\frac{\mu_2-\mu_1}{2}; +\infty)$

$$P_e = \frac{1}{\sqrt{2\pi}} \int_a^{+\infty} e^{-\frac{u^2}{2}} du.$$

(b) Use the inequality

$$\frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-(1/2)u^2} du \le \frac{1}{\sqrt{2\pi}a} e^{-(1/2)a^2}$$

to show that P_e goes to zero as $a = |\mu_2 - \mu_1|/\sigma$ goes to infinity.

Solution:

$$0 \le P_e \le \frac{1}{\sqrt{2\pi}\, a} e^{-\frac{a^2}{2}}$$

$$\lim_{a\to+\infty} 0 \le \lim_{a\to+\infty} P_e \le \lim_{a\to+\infty} \frac{e^{-\frac{a^2}{2}}}{\sqrt{2\pi}\, a} = 0$$

By squeezing theorem, $\lim_{a\to+\infty} P_e = 0$.

Problem 8. Recall that the probability mass function of a Poisson random variable is

$$P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!} \qquad x \in \{0, 1, \ldots, \infty\}$$

You are given two *equally likely* classes of Poisson data with parameters $\lambda_1 = 10$ and $\lambda_2 = 15$. This means that $x|\omega_1 \sim \text{Poisson}(\lambda_1)$ and $x|\omega_2 \sim \text{Poisson}(\lambda_2)$.

(a) Given the class conditionals, $x|\omega_1$ and $x|\omega_2$, find $P(\omega_1|x)$ in terms of λ_1, λ_2, $P(\omega_1)$, and $P(\omega_2)$. What type of function is the posterior?
Solution:

$$P_r(\omega_1|x) = \frac{P_r(x|\omega_1)\,P_r(\omega_1)}{P_r(x|\omega_1)\,P_r(\omega_1) + P_r(x|\omega_2)\,P_r(\omega_2)} = \frac{1}{1 + \frac{P_r(\omega_2)}{P_r(\omega_1)}\,\frac{e^{-\lambda_2}\lambda_2^x}{x!}\,\frac{x!}{e^{-\lambda_1}\lambda_1^x}}$$

$$= \frac{1}{1 + e^{(\ln P_r(\omega_2) - \ln P_r(\omega_1) - \lambda_2 + \lambda_1) + (\ln\lambda_2 - \ln\lambda_1)x}} = \frac{1}{1 + e^{-\theta x - \theta_0}}.$$

where $\theta = \ln\lambda_1 - \ln\lambda_2$.

$\theta_0 = \ln P_r(\omega_1) - \ln P_r(\omega_2) + \lambda_2 - \lambda_1$

The posterior is a logistic function.

(b) Find the optimal rule (decision boundary) for allocating an observation x to a particular class. Calculate the probability of correct classification for each class. Calculate the total error rate for this choice of decision boundary.
Solution:

Since the two classes are equally likely $Pr(w_1) = Pr(w_2) = \frac{1}{2}$.

$$Pr(w_1|x) = \frac{1}{1 + e^{-\theta x - \theta_0}} \qquad \theta = \ln\lambda_1 - \ln\lambda_2 = \ln\frac{2}{3} \qquad \theta_0 = 5$$

$$Pr(w_1|x) = \frac{1}{1 + 1.5^x \cdot e^{-5}} \qquad Pr(w_2|x) = 1 - Pr(w_1|x) = \frac{1.5^x e^{-5}}{1 + 1.5^x e^{-5}}$$

$$\exists x^* \quad Pr(w_1|x^*) = Pr(w_2|x^*) \quad \Rightarrow \quad 1.5^{x^*} = e^5 \quad x^* = \frac{5}{\ln 1.5} = 12.33$$

Thus for $x \in \{0, 1, \ldots 12\}$ x is classified in class 1.

for $x \in \{13, 14 \ldots\}$ x is classified in class 2.

For class 1, the probability of correct classification is

$$P_1 = \sum_{x=0}^{12} e^{-10}\frac{10^x}{x!} = 0.792.$$

For class 2, the probability of correct classification is

$$P_2 = \sum_{x=13}^{+\infty} e^{-15}\frac{15^x}{x!} = 1 - \sum_{x=0}^{12} e^{-15}\frac{15^x}{x!} = 0.732.$$

The error rate of this decision boundary is

$$P_e = P(w_1)(1 - P_1) + P(w_2)(1 - P_2) = 0.238.$$

(c) Suppose instead of one, we can obtain two independent measurements x_1 and x_2 for the object to be classified. How do the allocation rules and error rates change? Calculate the revised probability of correct classification for each class. Calculate the new total error in this case.

Hint: Always keep in mind that the Poisson distribution is defined for nonnegative integral values. Moreover, you can't be sure how much error you accumulate by erring on either side unless you explicitly calculate it.

Solution:

Similarly, the decision boundray can be computed using

$$Pr(w_1 \mid x_1, x_2) = Pr(w_2 \mid x_1, x_1).$$

$$\frac{Pr(x_1,x_2 \mid w_1)\, Pr(w_1)}{\sum_{k=1}^{2} Pr(x_1,x_2 \mid w_k)\, Pr(w_k)} = \frac{Pr(x_1,x_2 \mid w_2)\, Pr(w_2)}{\sum_{k=1}^{2} Pr(x_1,x_2 \mid w_k)\, P(w_k)} \Rightarrow Pr(x_1,x_2 \mid w_1) = Pr(x_1,x_2 \mid w_2)$$

Since x_1 and x_2 are independent measures.

$$Pr(x_1,x_2 \mid w_1) = \frac{e^{-20}\, 10^{x_1+x_2}}{x_1! \cdot x_2!}, \quad Pr(x_1,x_2 \mid w_2) = \frac{e^{-30}\, 15^{x_1+x_2}}{x_1!\, x_2!}$$

Let $\bar{x} = x_1 + x_2$. $\quad 1.5^{\bar{x}} = e^{10}$. $\quad \bar{x} = 24.663$.

Thus. if $x_1+x_2 \in \{0, \ldots 24\}$ it should be classified in class 1.

if $x_1+x_2 \in \{25, 26, \ldots\}$ it should be classfied in class 2.

For class 1, The probability of correct classification is.

$$P_1 = \sum_{n=0}^{24} \sum_{k=0}^{n} \frac{e^{-20}\, 10^{n}}{k!\,(n-k)!} = 0.8432.$$

For class 2, the probability of correct classiftcation is

$$P_2 = 1 - \sum_{n=0}^{24} \sum_{k=0}^{n} \frac{e^{-30} 15^{n}}{k!\,(n-k)!} = 0.8428.$$

The error rate is

$$P_e = Pr(w_1)(1-P_1) + Pr(w_2)(1-P_2) = 0.1570.$$

With an additional measurement, the error rate is decreased.

That's because for some extreme cases such as $x_1 = 0$ $x_2 = 13$ with to measurements it is regared as in class one but with $x = 13$, it will be regarded as in class 2 which is probably wrong.

Problem 9 (Optional: Extra for Experts) . Let $X_1, X_2, ..., X_n$ be a sequence of points chosen independently and uniformly from within a 2-dimensional unit ball $B = \{x \in \mathbb{R}^2 : {x_1}^2 + {x_2}^2 \leq 1\}$. A set of points $X_1, X_2, ...X_n$ lie in a hemisphere if there is a line passing through the origin for which all n points lie on a particular side of the hemisphere. Define the event:

$$A_n = \{X_1, X_2 ... X_n \text{ lie in a hemisphere}\}$$

Compute $\Pr\{A_n\}$. (There are multiple ways of doing this. Some are simpler than others)

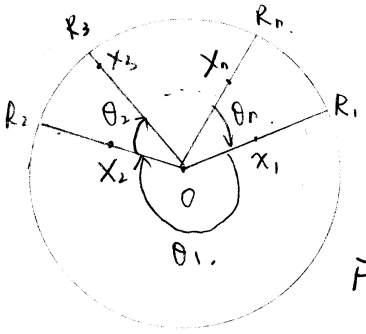
Credit and Thanks to Professor Thomas Courtade for writing this question

Solution: Without loss of generality, let's assume the points are numbered in clockwise manner.

Notations : O : original point

R_i : connect O and X_i, R_i is the cross over point on the circle. such that X_i is on radius OR_i.

θ_i : the angle from OR_i to OR_{i+1}. for $i < n$. θ_n is the angle from OR_n to OR_1.



Facts : ① The probability of X_i on original is zero.

② The probability of X_j on OR_i, $i \neq j$, is zero.

A_n is equivalent to one of θ_i is greater than π.

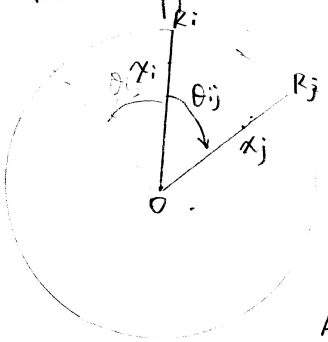
$$\Pr(A_n) = \sum_{i=1}^{n} \Pr(\theta_i > \pi) = n \Pr(\theta_1 > \pi)$$

$\theta_1 > \pi$ is equivalent to $X_2, \cdots X_n$ in a hemisphere defined by $O X_1 R_1$.

$$\Pr(\theta_1 > \pi) = \frac{1}{2^{n-1}}$$

Thus. $$\Pr(A_n) = \frac{n}{2^{n-1}}$$

Another approach.



θ_{ij} is defined as depicted in the left graph. (θ_{ij} is the smaller angle)

θ_{ij} is uniformly distributed on $[0, \pi]$

$$\Pr(\theta_{ij} \leq \theta) = \frac{2 \cdot \text{Area}(OR_1R_2)}{\text{Area(circle)}}$$ since X_i uniformly distributed in B

$$\Pr(\theta_{ij} \leq \theta) = \frac{\frac{\theta}{2} \cdot 2}{\pi} = \frac{\theta}{\pi}$$

A_n is equivalent to pick i from $\{1, \ldots n\}$. $\theta_{ij} < \pi$. $\forall j \neq i$.

$$\Pr(A_n) = \binom{n}{1} \prod_{\substack{j=1 \\ j \neq i}}^{n} \Pr(\theta_{ij} < \pi) = n \cdot \frac{1}{2^{n-1}} = \frac{n}{2^{n-1}}$$