CS189: Introduction to Machine Learning

Homework 2 Solutions

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: The expected value is

$$\begin{aligned} &\int_0^{1/\sqrt{3}} 4\frac{2}{\pi(1+x^2)}dx + \int_{1/\sqrt{3}}^{1} 3\frac{2}{\pi(1+x^2)}dx + \int_1^{\sqrt{3}} 2\frac{2}{\pi(1+x^2)}dx \\ &= \frac{2}{\pi}\left[4\left(\tan^{-1}\frac{1}{\sqrt{3}} - \tan^{-1}0\right) + 3\left(\tan^{-1}1 - \tan^{-1}\frac{1}{\sqrt{3}}\right) + 2\left(\tan^{-1}\sqrt{3} - \tan^{-1}1\right)\right] \\ &= \boxed{\frac{13}{6}} \end{aligned}$$

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 $x_5 = 2.4$.

Solution: We'll solve the general case for the MLE of an exponential distribution, then plug in the numbers we have.

$$P(X_1, X_2, \ldots, X_n|\theta) = P(X_1|\theta)P(X_2|\theta)\ldots P(X_n|\theta)$$

$$L(\theta|X_1, X_2, \ldots, X_n) = P(X_1|\theta)P(X_2|\theta)\ldots P(X_n|\theta)$$

$$L(\theta|X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\theta) = \theta^n \exp(-\theta\sum_{i=1}^{n} X_i)$$

Finding the log-likelihood:

$$l(\theta|X_1, X_2, \ldots, X_n) = n\ln(\theta) - \theta\sum_{i=1}^{n} X_i$$

Taking the derivative with respect to θ:

$$\frac{\delta l}{\delta \theta} = \frac{n}{\theta} - \sum_{i=1}^{n} X_i = 0$$

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} X_i}$$

Plugging in our values for X_i, we get $\hat{\theta} = 0.88$.

Problem 3. The polynomial kernel is defined to be

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

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.

Solution: (from Wikipedia)

$$K(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{n} x_i y_i + c\right)^2 = \sum_{i=1}^{n} \left(x_i^2\right)\left(y_i^2\right) + \sum_{i=2}^{n}\sum_{j=1}^{i-1} \left(\sqrt{2}x_i x_j\right)\left(\sqrt{2}y_i y_j\right) + \sum_{i=1}^{n} \left(\sqrt{2c}x_i\right)\left(\sqrt{2c}y_i\right) + c^2$$

$$\Phi(\mathbf{z}) = \langle x_n^2, \ldots, x_1^2, \sqrt{2}x_n x_{n-1}, \ldots, \sqrt{2}x_n x_1, \sqrt{2}x_{n-1}x_{n-2}, \ldots, \sqrt{2}x_{n-1}x_1, \ldots, \sqrt{2}x_2 x_1, \sqrt{2c}x_n, \ldots, \sqrt{2c}x_1, c\rangle$$

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

Solution: Cross Validation

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

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 $x^\top A x$. Write your answer as a sum involving the elements of A and x.

Solution:

$$x^\top A x = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$$

(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: Let $i \in [1, n]$, and let e_i be the i^{th} standard basis vector (that is, the vector of all zeros except for a single 1 in the i^{th} position). Then, by the positive definitness of A, we have $e_i^\top A e_i = a_{ii} > 0$.

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

Solution: Let $x \neq 0$. Then

$$\begin{aligned} x^\top (B + \gamma I) x &= x^\top B x + x^\top \gamma I x \\ &= x^\top B x + \gamma \|x\|^2 \\ &> 0 \end{aligned}$$

because $x^\top B x \geq 0$ (since B is positive semidefinite) and $\|x\|^2 > 0$ (because $x \neq 0$). Hence $B + \gamma I$ is positive definite.

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.

(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: Define $\lambda_{ij} = \ell(f(x) = i, y = j)$. The risk of classifying a new datapoint as class i is

$$R(\alpha_i|x) = \sum_j \lambda_{ij} P(\omega_j|x) = \lambda_s(1 - P(\omega_i|x)),$$

and the risk of classifying the new datapoint as doubt is

$$R(\alpha_{c+1}|x) = \lambda_r \sum_j P(\omega_j|x) = \lambda_r.$$

For choosing doubt to be better than choosing any of the classes, the ratio of the risks must satisfy

$$1 > \frac{R(\alpha_{c+1}|x)}{R(\alpha_i|x)} = \frac{\lambda_r}{\lambda_s(1 - P(\omega_i|x))} \implies P(\omega_i|x) < 1 - \frac{\lambda_r}{\lambda_s}$$

for all i. This means that any particular i for which $P(\omega_i|x) \geq 1 - \frac{\lambda_r}{\lambda_s}$ should not be assigned doubt. In this case, the class to choose must be

$$\arg \min_{1 \leq i \leq c} R(\alpha_i|x) = \arg \min_{1 \leq i \leq c} \lambda_s(1 - P(\omega_i|x)) = \arg \max_{1 \leq i \leq c} P(\omega_i|x),$$

as required.

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

Solution: If $\lambda_r = 0$, then doubt will always be assigned, since for all i, $P(\omega_i|x) \geq 1 - \lambda_r/\lambda_s = 1$ is not satisfied unless $P(\omega_i|x) = 1$.

If $\lambda_r > \lambda_s$, then doubt will never be assigned, since for all i, $P(\omega_i|x) \geq 0 > 1 - \lambda_r/\lambda_s$ always holds.

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: Without loss of generality, assume $\mu_1 \leq \mu_2$.

Since the variances of the class conditionals are equal, we will have a linear decision boundary, so all we need to find is the point x_0 at which $P(\omega_1|x) = P(\omega_2|x)$. This point must satisfy, by Bayes' rule,

$$\begin{aligned} & P(x_0|\omega_1)P(\omega_1) = P(x_0|\omega_2)P(\omega_2) \\ \implies \quad & \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\frac{(x_0-\mu_1)^2}{\sigma^2}\right] = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\frac{(x_0-\mu_2)^2}{\sigma^2}\right] \\ \implies \quad & x_0 = \frac{\mu_1+\mu_2}{2}. \end{aligned}$$

Then

$$\begin{aligned} P_e &= \int_{-\infty}^{\infty} P(\text{error}|x)P(x)dx \\ &= \int_{-\infty}^{x_0} P(\omega_2|x)P(x)dx + \int_{x_0}^{\infty} P(\omega_1|x)P(x)dx \\ &= \frac{1}{2}\int_{-\infty}^{x_0} P(x|\omega_2)dx + \frac{1}{2}\int_{x_0}^{\infty} P(x|\omega_1)dx \end{aligned}$$

Since the integrands are symmetric around x_0, we have

$$\begin{aligned} &= 2\left[\frac{1}{2}\int_{x_0}^{\infty} P(x|\omega_1)dx\right] \\ &= \int_{\frac{\mu_1+\mu_2}{2}}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\frac{(x-\mu_1)^2}{\sigma^2}\right] dx \end{aligned}$$

Substituting $u = \frac{x-\mu_1}{\sigma}$ gives

$$= \int_a^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) du$$

where $a = \frac{\mu_2-\mu_1}{2\sigma}$. (If we relax the $\mu_1 \leq \mu_2$ assumption, we can instead write $a = \frac{|\mu_2-\mu_1|}{2\sigma}$.)

(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: Applying the bound shows that as a goes to infinity,

$$P_e = \int_a^\infty \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)du \le \frac{1}{\sqrt{2\pi}a}\exp\left(-\frac{1}{2}a^2\right) = \frac{1}{\sqrt{2\pi}a\exp\left(\frac{1}{2}a^2\right)}$$

goes to 0, because the denominator goes to infinity. This means that we can get lower error as the class conditional distributions move farther apart.

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: We use Bayes' Rule

$$\begin{aligned}
P(\omega_1|x) &= \frac{P(x|\omega_1)P(\omega_1)}{P(x|\omega_1)P(\omega_1) + P(x|\omega_2)P(\omega_2)} \\
&= \frac{e^{-\lambda_1}\frac{\lambda_1^x}{x!}P(\omega_1)}{e^{-\lambda_1}\frac{\lambda_1^x}{x!}P(\omega_1) + e^{-\lambda_2}\frac{\lambda_2^x}{x!}P(\omega_2)} \\
&= \frac{1}{1 + e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^x\frac{P(\omega_2)}{P(\omega_1)}} \\
&= \frac{1}{1 + e^{-(\lambda_1+\lambda_2)+\ln\frac{P(\omega_2)}{P(\omega_1)}+x\ln(\lambda_1+\lambda_2)}}
\end{aligned}$$

The posterior is a logistic function. If $P(\omega_1) = P(\omega_2)$, then

$$P(\omega_1|x) = \frac{1}{1 + e^{-(\lambda_1+\lambda_2)+x\ln(\lambda_1+\lambda_2)}}$$

We accepted either answer

(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:

The decision boundary is the value for x for which

$$1 < \frac{P(\omega_1|x)}{P(\omega_2|x)} = e^{\lambda_2 - \lambda_1} \left(\frac{\lambda_1}{\lambda_2}\right)^x$$

so we should choose class 1 if

$$\boxed{x < \frac{\lambda_1 - \lambda_2}{\ln \lambda_1 - \ln \lambda_2} \approx 12.3}$$

or class 2 otherwise (the inequality holds since $\lambda_1/\lambda_2 < 1$).

Recall that the probability of correctly classifying a vector $\vec{x}$ is $1 - P(error) = 1 - \sum_{-\infty}^{\infty} P(error|x)P(x)$. For the decision boundary $\theta = 12.3$, it would therefore equal:

$$\sum_{i=0}^{12} P(\omega_1|x)P(x) + \sum_{i=13}^{\infty} P(\omega_2|x)P(x) = \sum_{i=0}^{12} P(x|\omega_1)P(w_1) + \sum_{i=13}^{\infty} P(x|\omega_2)P(\omega_2)$$

.

The probability of correctly classifying as class 1 is

$$P(x < 12.3|\omega_1)P(\omega_1) = \sum_{x=0}^{12} e^{-10}10^x/x! * 0.5 \approx \boxed{0.396}$$

and the probability of correctly classifying as class 2 is

$$P(x > 12.3|\omega_2)P(\omega_2) = (1 - \sum_{x=0}^{12} e^{-15}15^x/x!) * 0.5 \approx \boxed{0.366}$$

The total error rate for this decision boundary is $1 - 0.396 - 0.366 \approx 0.238$.

(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.

Solution: If we receive two independent measurements x_1, x_2, then the decision boundary condition for choosing class 1 is

$$1 < \frac{P(\omega_1|x_1, x_2)}{P(\omega_2|x_1, x_2)} = \frac{P(x_1, x_2|\omega_1)P(\omega_1)}{P(x_1, x_2|\omega_2)P(\omega_2)} = e^{2(\lambda_2 - \lambda_1)} \left(\frac{\lambda_1}{\lambda_2}\right)^{x_1 + x_2}$$

which means that we should choose class 1 if

$$\boxed{x_1 + x_2 < 2\frac{\lambda_1 - \lambda_2}{\ln \lambda_1 - \ln \lambda_2} \approx 24.6}$$

The probability of correctly classifying as class 1 is

$$P(x_1 + x_2 < 24.6|\omega_1)P(\omega_1) = \sum_{x=0}^{24} e^{-20}20^x/x! * 0.5 \approx \boxed{0.4216}$$

(since $x_1 + x_2 \sim$ Poisson$(2\lambda_1)$ assuming x_1, x_2 are iid from class 1), and the probability of correctly classifying as class 2 is

$$P(x_1 + x_2 > 24.6|\omega_2)P(\omega_2) = 1 - \sum_{x=0}^{24} e^{-30}30^x/x!0.5 \approx \boxed{0.4214}$$

(since $x_1 + x_2 \sim$ Poisson$(2\lambda_2)$ assuming x_1, x_2 are iid from class 2). The total error rate for this decision boundary is

$$1 - 0.4216 - 0.4214 \approx \boxed{0.157}$$

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.

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

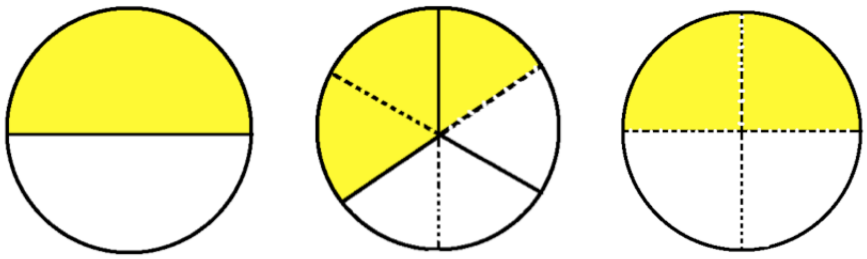
Solution: Professor Courtade's Solution (Projection)

We are asked to evaluate the probability that n randomly chosen points lie in a hemisphere. Let us call the points $\{X_1, X_2, ...X_n\}$. For $1 \leq i \leq n$, let the diametrically opposite point of X_i be $\hat{X}_i$. Construct a radius of the circle which contains the segment $(0, X_i)$, and rotate this radius anticlockwise, until it hits the point $\hat{X}_i$.Define the hemisphere swept by this radius as S_i. The event "$\{X_1, X_2, ...X_n\}$ lie in a hemisphere" is a union of the disjoint events E_i, where E_i is defined as $\{X_1, X_2...X_n\}$g lie in the hemisphere S_i". Since X_i's are i.i.d., we immediately obtain $Pr(E_i) = \frac{1}{2^{n-1}}$, and conclude that the probability of all n points lying in a hemisphere is $\frac{n}{2^{n-1}}$.

Solution: Daniel Xu's Solution (Inclusion-Exclusion)

Consider a discretized version of the problem, where we have parameters

$$M = \#\text{ of equally spaced hemisphere sections} \quad (1)$$
$$N = \#\text{ of points } X_1, X_2, ...X_n \quad (2)$$



= 1 Particular Hemisphere Section (M in total)

Thus, we have

$$P(A_{n,m}) = P(X_1, X_2, ...X_n \text{ lie in one of the } H_m \text{ hemisphere sections}) \tag{3}$$
$$= M((\frac{1}{2})^N - (\frac{1}{2} - \frac{1}{M})^N) \tag{4}$$

Why? Since the set of points $\{X_1, X_2, ...X_n\}$ are uniformly distributed in the 2D sphere, i.i.d, and any particular hemisphere has an area of $\frac{1}{2}$ the 2D ball, the probability of the set of points falling in a particular hemisphere H_i is as follows:

$$P(X_1, X_2...X_n \text{ lie in a particular hemisphere } H_i) = (\frac{1}{2})^N$$

Since hemispheres H_i $\forall i \in [0...M]$ are identical to each other, the points have an equal chance of falling in any of the M hemisphere sections. Thus,

$$P(A_{n,m,i}) = M * P(X_1, X_2, ...X_N \text{ lie in a particular hemisphere } H_i)$$

However, notice that when $M > 2$, these hemispheres will overlap with each other. Thus, by the Inclusion-Exclusion Principle, we must subtract the overlap probability mass:

$$\text{Overlap Probability Mass} = (\frac{1}{2} - \frac{1}{M})^N$$

There are M of these overlaps, and each are of the same size. Thus, we have:

$$P(A_{n,m}) = M * P(A_{n,m,i}) - M * (\text{Overlap Probability Mass}) \tag{5}$$
$$= M(\frac{1}{2})^N - M(\frac{1}{2} - \frac{1}{M})^N \tag{6}$$
$$= M((\frac{1}{2})^N - (\frac{1}{2} - \frac{1}{M})^N) \tag{7}$$

Thus,

$$P(A_{n,m}) = \frac{M}{2^N}(1 - (\frac{M-2}{M})^N) \tag{8}$$
$$= \frac{M}{2^N}(1^N - (\frac{M-2}{M})^N) \tag{9}$$
$$= \frac{M}{2^N}((1 - \frac{M-2}{M})(\sum_{i=0}^{N-1}(\frac{M-2}{M})^i * 1^{N-1-i})) \quad = \frac{1}{2^N} * 2 * (\sum_{i=0}^{N-1}(1 - \frac{2}{M})^i)) \tag{10}$$

The second to last step comes from the Sum/Difference of Two Nth Powers identity (a proof of it is available on Wikipedia):

$$a^N - b^N = (a-b)(\sum_{i=0}^{N-1} b^i a^{N-1-i}) \tag{11}$$
$$\text{where } a = 1 \text{ , } b = \frac{M-2}{M} \tag{12}$$

Now, to get the continuous version of this problem in order to find the probability that the set of points $X_1, X_2, ...X_n$ fall in any possible hemisphere of the 2D ball, we take the limit as the number of discrete hemispheres go to infinity.

$$\lim_{M\to\infty} P(A_{n,m}) = \frac{1}{2^N}(2)\sum_{i=0}^{N-1} *1^i = \frac{N}{2^{N-1}}$$