CS 189 Spring 2022 | Introduction to Machine Learning Jonathan Shewchuk

Exam Prep 1

This exam-prep discussion section covers Bayesian decision theory and maximum likelihood estimation. In order, the questions were taken from the Spring offerings in 2016, 2016, 2017, 2019, 2020, 2019, and 2017.

1 Multiple Choice

(f) [3 pts] The Bayes risk for a decision problem is zero when

○ the class distributions $P(X|Y)$ do not overlap.

○ the training data is linearly separable.

○ the loss function $L(z, y)$ is symmetrical.

○ the Bayes decision rule perfectly classifies the training data.

(g) [3 pts] Let $L(z, y)$ be a loss function (where y is the true class and z is the predicted class). Which of the following loss functions will *always* lead to the same Bayes decision rule as L?

○ $L_1(z, y) = aL(z, y)$, $a > 0$

○ $L_2(z, y) = aL(z, y)$, $a < 0$

○ $L_3(z, y) = L(z, y) + b$, $b > 0$

○ $L_4(z, y) = L(z, y) + b$, $b < 0$

(t) [3 pts] Which of the following statements about maximum likelihood estimation are true?

○ MLE, applied to estimate the mean parameter μ of a normal distribution $\mathcal{N}(\mu, \Sigma)$ with a known covariance matrix Σ, returns the mean of the sample points

○ MLE, applied to estimate the covariance parameter Σ of a normal distribution $\mathcal{N}(\mu, \Sigma)$, returns $\hat{\Sigma} = \frac{1}{n}X^T X$, where X is the design matrix

○ For a sample drawn from a normal distribution, the likelihood $\mathcal{L}(\mu, \sigma; X_1, \ldots, X_n)$ is equal to the probability of drawing exactly the points $X_1, \ldots, X_n$ (in that order) when you draw n random points from $\mathcal{N}(\mu, \sigma)$

○ Maximizing the log likelihood is equivalent to maximizing the likelihood

(s) [3 pts] Suppose you have a sample in which each point has d features and comes from class C or class D. The class conditional distributions are $(X_i|y_i = C) \sim N(\mu_C, \sigma_C^2)$ and $(X_i|y_i = D) \sim N(\mu_D, \sigma_D^2)$ for unknown values $\mu_C, \mu_D \in \mathbb{R}^d$ and $\sigma_C^2, \sigma_D^2 \in \mathbb{R}$. The class priors are π_C and π_D. We use 0-1 loss.

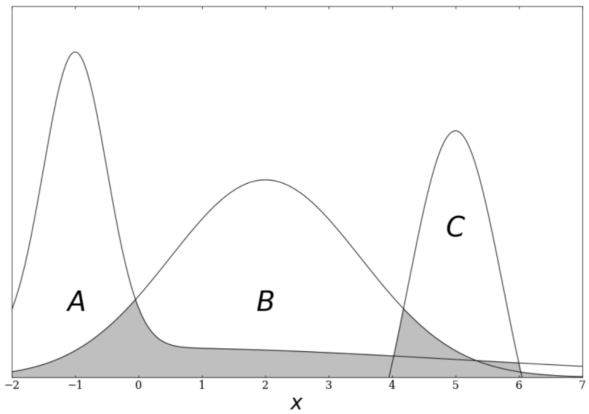
○ If $\pi_C = \pi_D$ and $\sigma_C = \sigma_D$, then the Bayes decision rule assigns a test point z to the class whose mean is closest to z.

○ If $\pi_C = \pi_D$, then the Bayes decision rule is $r^*(z) = \text{argmin}_{A \in \{C,D\}} \left(|z - \mu_A|^2/(2\sigma_A^2) + d \ln \sigma_A\right)$

○ If $\sigma_C = \sigma_D$, then the Bayes decision boundary is always linear.

○ If $\sigma_C = \sigma_D$, then QDA will always produce a linear decision boundary when you fit it to your sample.

(j) [4 pts] The following chart depicts the class-conditional distributions $P(X|Y)$ for a classification problem with three classes, A, B, and C. Classes A and B are normally distributed over the domain $(-\infty, \infty)$; Class C is defined only over the finite domain depicted below. All three classes have prior probabilities π_A, π_B, π_C **strictly greater than zero**; the chart does **not** show the influence of these priors. We use the **0-1 loss** function.



○ A: The Bayes risk is the area of the shaded region in the chart (including the area not depicted off the sides of the chart, going to $x = \pm\infty$)

○ B: Depending on the priors, it is possible that the Bayes rule $r^*(x)$ will classify all inputs as class B

○ C: Depending on the priors, it is possible that the Bayes rule $r^*(x)$ will classify all inputs as class C

○ D: Depending on the priors, it is possible that the Bayes risk is zero

2 Free Response

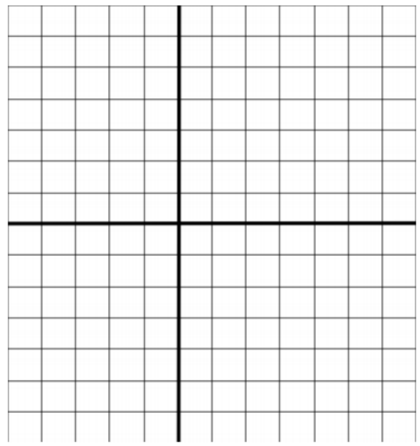
Q3. [10 pts] Quadratic Discriminant Analysis

(a) [4 pts] Consider 12 labeled data points sampled from three distinct classes:

Class 0: $\begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} -2 \\ 0 \end{bmatrix}, \begin{bmatrix} 5 \\ 3 \end{bmatrix}, \begin{bmatrix} -3 \\ -5 \end{bmatrix}$ Class 1: $\begin{bmatrix} \sqrt{2} \\ \sqrt{2} \end{bmatrix}, \begin{bmatrix} -\sqrt{2} \\ \sqrt{2} \end{bmatrix}, \begin{bmatrix} 4\sqrt{2} \\ -\sqrt{2} \end{bmatrix}, \begin{bmatrix} -4\sqrt{2} \\ -\sqrt{2} \end{bmatrix}$ Class 2: $\begin{bmatrix} 3 \\ 5 \end{bmatrix}, \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \begin{bmatrix} 8 \\ 6 \end{bmatrix}, \begin{bmatrix} 0 \\ -2 \end{bmatrix}$

For each class $C \in \{0, 1, 2\}$, compute the class sample mean μ_C, the class sample covariance matrix Σ_C, and the estimate of the prior probability π_C that a point belongs to class C. (Hint: $\mu_1 = \mu_0$ and $\Sigma_2 = \Sigma_0$.)

(b) [4 pts] Sketch one or more isocontours of the QDA-produced normal distribution or quadratic discriminant function (they each have the same contours) for each class. The isovalues are not important; the important aspects are the centers, axis directions, and relative axis lengths of the isocontours. Clearly label the centers of the isocontours and to which class they correspond.



(c) [2 pts] Suppose that we apply LDA to classify the data given in part (a). Why will this give a poor decision boundary?

Q3. [10 pts] Maximum Likelihood Estimation for Reliability Testing

Suppose we are reliability testing n units taken randomly from a population of identical appliances. We want to estimate the mean failure time of the population. We assume the failure times come from an exponential distribution with parameter $\lambda > 0$, whose probability density function is $f(x) = \lambda e^{-\lambda x}$ (on the domain $x \geq 0$) and whose cumulative distribution function is $F(x) = \int_0^x f(x)\,\mathrm{d}x = 1 - e^{-\lambda x}$.

(a) [6 pts] In an ideal (but impractical) scenario, we run the units until they all fail. The failure times are $t_1, t_2, \ldots, t_n$.

Formulate the likelihood function $\mathcal{L}(\lambda; t_1, \ldots, t_n)$ for our data. Then find the maximum likelihood estimate $\hat{\lambda}$ for the distribution's parameter.

(b) [4 pts] In a more realistic scenario, we run the units for a fixed time T. We observe r unit failures, where $0 \leq r \leq n$, and there are $n - r$ units that survive the entire time T without failing. The failure times are $t_1, t_2, \ldots, t_r$.

Formulate the likelihood function $\mathcal{L}(\lambda; n, r, t_1, \ldots, t_r)$ for our data. Then find the maximum likelihood estimate $\hat{\lambda}$ for the distribution's parameter.

Hint 1: What is the probability that a unit will not fail during time T? *Hint 2:* It is okay to define $\mathcal{L}(\lambda)$ in a way that includes contributions (densities and probability masses) that are not commensurate with each other. Then the constant of proportionality of $\mathcal{L}(\lambda)$ is meaningless, but that constant is irrelevant for finding the best-fit parameter $\hat{\lambda}$. *Hint 3:* If you're confused, for part marks write down the likelihood that r units fail and $n-r$ units survive; then try the full problem. *Hint 4:* If you do it right, $\hat{\lambda}$ will be the number of observed failures divided by the sum of unit test times.