Instructor Yusuf Yaslan
Assistant Kıymet Kaya

- **Please do your homework on your own.** You are encouraged to discuss the questions with your class mates, but the code and the homework you submitted must be your own work. Cheating is highly discouraged for it could mean a zero or negative grade from the homework.
- **Late submissions will not be accepted**. Please do not email us your late submissions.
- Unless we indicate otherwise, **do not** use libraries for machine learning methods. When in doubt, or if you have a question related to the homework, reach us via email.
- Submissions are expected to include a pdf file prepared with Latex that contains the solutions and a Jupyter Notebook file with Python for the coding-related questions.

Please read the instructions below for coding-related questions

- Install a Conda environment if you do not have it already.
- Install Jupyter Notebook.
- In this homework, you are expected to use matplotlib and numpy.
- Questions or sections of a question that are marked with red color (e.g. a), 5)) should be done in Jupyter Notebook.
- Do not forget to format your code and leave comments for non-trivial sections.

Make sure that you read chapter 4-5 and Appendix A from the textbook.

Question 1

(a) Let

$$f_X(x) = \begin{cases} x, & 0 \leq x \leq 1 \\ -x, & -1 \leq x < 0 \\ 0, & \text{otherwise} \end{cases}$$

and $Y = X^2$.

(i) Find $\rho(X, Y)$ (ρ: correlation coefficient).
(ii) Are X and Y independent?

(b) Let $X \sim \mathcal{N}(0, 1)$ and $Y = WX$, where $p(W = -1) = p(W = 1) = 0.5$.

(i) Find the marginal density function of Y.
(ii) Find $cov[X, Y]$.
(iii) Are X and Y independent?

(c) Let $f_{XY}(X, Y)$ be density function of bivariate Gaussian.

(i) Write the joint distribution of X, Y.
(ii) When is X and Y independent?

Question 2

In a toy manufacturing company, parts of the toys are produced by a plastic molding machine with a success rate of θ. After a batch of molding of size K, L many parts are faulty, and the remaining $K - L$ many parts are successfully produced.

(a) We believe that there are two types of molding machines and put our prior knowledge as follows:

$$p(\theta) = \begin{cases} 0.5, & \theta = 0.4 \\ 0.5, & \theta = 0.6 \end{cases}$$

Given the prior, find the MAP estimate of the parameter θ in terms of K and L.

(b) We can model the prior knowledge with Beta distribution $p(\theta) = \text{Beta}(\theta; \alpha, \beta)$. Find the MAP estimator $\hat{\theta}$ with this prior.

(c) Explain why we consider Beta distribution as conjugate prior for the likelihood function in this problem.

Question 3

Suppose for a classification problem with C classes and F features, we want to fit a naive Bayes classifier. All the features are binary, $x_j \in 0, 1$, class conditional density is represented with $p(x|y)$, parameters are denoted as θ_{jc} for $p(x_j = 1 \mid y = c)$ where j denotes the feature id and y is the class label for class $c \in C$.

(a) Under the naive Bayes assumption, give the formula for $p(\boldsymbol{x} \mid y = c)$ and state how many parameters are required.

(b) If we drop the naive Bayes assumption and consider a model such that no factorization assumptions are made (i.e. features are not independent), give the formula for $p(\boldsymbol{x} \mid y = c)$ and state how many parameters are required.

(c) Suppose F (number of features) is a fixed value and we have T training samples. For the cases below, state which model (i.e. a) Naive Bayes b) no factorization) would be more likely to result in lower test set error.

(i) T is very small.

(ii) T is very large.

Question 4

As seen in the first homework, Kullback-Leibler (KL) divergence between two distributions $p(x)$ and $q(x)$ is given by

$$KL(p||q) = -\int p(x) \ln \frac{q(x)}{p(x)} dx$$

In some machine learning applications, the computational cost for learning and inference can be too high if there are too many features. Moreover, too many features can cause overfitting in Naive Bayes Classifiers. 1 way to handle this issue is by selecting a subset of features according to their relevance to the classification problem. One method to estimate the relevance of a feature is calculating its mutual information with the class labels and then selecting the top K features with the highest mutual information.
For discrete distributions, mutual information is defined as:

$$\begin{aligned} I(X;Y) &= D_{KL}(P_{(X,Y)}||P_X \otimes P_Y) \\ &= \sum_{y \in Y} \sum_{x \in X} p_{(X,Y)}(x,y) \log(\frac{p_{(X,Y)}(x,y)}{p_X(x) p_Y(y)}) \end{aligned}$$

Suppose naive Bayes classifiers are used and feature selection will be applied in order to select the top K features out of N features. Mutual information of each feature with the class labels will be calculated in order to determine a relevance measure.
Derive the mutual information formula for the following cases:

(a) All the features are binary.

(b) All the features are categorical with M categories.

Question 5

Design a classifier with the data given within the homework folder. Divide the data into train and evaluation sets with three to one ratio respectively. Assume that the class-conditional densities are Multivariate Normal.

a) Design a linear discriminant classifier with a shared covariance matrix using MLE estimators. Plot a contour plot that shows the decision boundary. Similar to Figure 1.

b) Design a quadratic discriminant classifier using distinct covariance matrices and obtain the training and evaluation accuracy. Plot a contour plot that shows the decision boundary.

c) Now, assume that we know the prior distributions of the mean parameters such that $\theta_1 \sim \mathcal{N}([0,0]^T, \sigma I)$ and $\theta_2 \sim \mathcal{N}([4,0]^T, \sigma I)$ where σ=0.5 for the first and second class respectively. Instead of MLE use the MAP estimator for the mean parameter and analytically derive the estimator (use S for the covariance matrix and $\bar{\mu}$ for the mean vector of the data).

d) Repeat a) and b) with the new estimator.

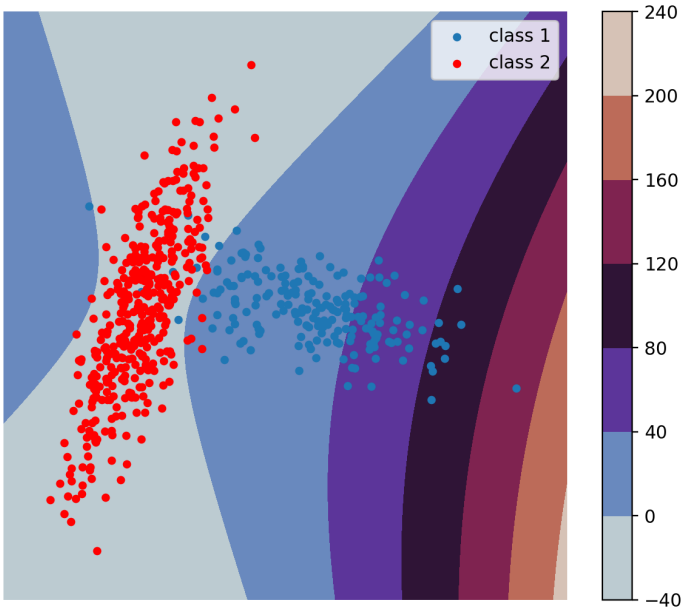

Figure 1: Example quadratic decision boundary plot