1. Assune that in a c-class classification problem, we have k features $X_1, X_2, \ldots, X_k$ that are independent conditioned on the class label and $X_j|\omega_i \sim \text{Gamma}(p_i, \lambda_j)$, i.e. $p_{X_j|\omega_i}(x_j|\omega_i) = \frac{1}{\Gamma(p_i)}\lambda_j^{p_i} x_j^{p_i-1} e^{-\lambda_j x_j}, p_i, \lambda_j > 0$. (30 pts)

   (a) Determine the Bayes' optimal classifier's decision rule making the general assumption that the prior probability of the classes are different.

   (b) When are the decision boundaries linear functions of $x_1, x_2, \ldots, x_k$?

   (c) Assuming that $p_1 = 4, p_2 = 2, c = 2, k = 4, \lambda_1 = \lambda_3 = 1, \lambda_2 = \lambda_4 = 2$, and that the prior probabilites of each class are equal, classify $\mathbf{x} = (0.1, 0.2, 0.3, 4)$.

   (d) Assuming that $p_1 = 3.2, p_2 = 8, c = 2, k = 1, \lambda_1 = 1$, and that the prior probabilites of each class are equal, find the decision boundary $x = x^*$. Also, find the probability of type-1 and type-2 errors.

   (e) Assuming that $p_1 = p_2 = 4, c = 2, k = 2, \lambda_1 = 8, \lambda_2 = 0.3$, and $P(\omega_1) = 1/4, P(\omega_2) = 3/4$, find the decision boundary $f(x_1, x_2) = 0$.

2. Assume that in a c-class classification problem, there are k conditionally independent features and $X_i|\omega_j \sim \text{Lap}(m_{ij}, \lambda_i)$, i.e. $p_{X_i|\omega_j}(x_i|\omega_j) = \frac{\lambda_i}{2} e^{-\lambda_i|x_i - m_{ij}|}, \lambda_i > 0, i \in \{1, 2, \ldots, k\}, j \in \{1, 2, \ldots, c\}$. Assuming that the prior class probabilities are equal, show that the minimum error rate classifier is also a minimum weighted Manhattan distance (or weighted $\mathscr{L}_1$-distance) classifier. When does the minimum error rate classifier becomes the minimum Manhattan distance classifier? (15 pts)

3. The class-conditional density functions of a discrete random variable X for four pattern classes are shown below: (20 pts)

x	$p(x\|\omega_1)$	$p(x\|\omega_2)$	$p(x\|\omega_3)$	$p(x\|\omega_4)$
1	1/3	1/2	1/6	2/5
2	1/3	1/4	1/3	2/5
3	1/3	1/4	1/2	1/5

   The loss function $\lambda(\alpha_i|\omega_j)$ is summarized in the following table, where action α_i means decide pattern class ω_i:

	ω_1	ω_2	ω_3	ω_4
α_1	0	2	3	4
α_2	1	0	1	8
α_3	3	2	0	2
α_4	5	3	1	0

   Assume $P(\omega_1) = 1/10, P(\omega_2) = 1/5, P(\omega_3) = 1/2, P(\omega_4) = 1/5$.

   (a) Compute the conditional risk for each action as:
   $R(\alpha_i|x) = \sum_{j=1}^{4} \lambda(\alpha_i|\omega_j) p(\omega_j|x)$

(b) Compute the overall risk R as:
$R = \sum_{i=1}^{3} R(\alpha(x_i)|x_i)p(x_i)$
where $\alpha(x_i)$ is the decision rule minimizing the conditional risk for x_i.

4. The following data set was collected to classify people who evade taxes:

Tax ID	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	122 K	No
2	No	Married	77 K	No
3	No	Married	106 K	No
4	No	Single	88 K	Yes
5	Yes	Divorced	210 K	No
6	No	Single	72 K	No
7	Yes	Married	117 K	No
8	No	Married	60 K	No
9	No	Divorced	90 K	Yes
10	No	Single	85 K	Yes

Considering relevant features in the table (only one feature is not relevant), assume that the features are *conditionally independent*. (25 pts)

(a) Estimate prior class probabilities.

(b) For continuous feature(s), assume conditional Gaussianity and estimate class conditional pdfs $p(x|\omega_i)$. Use Maximum Likelihood Estimates.

(c) For each discrete feature X, assume that the number of instances in class ω_i for which $X = x_j$ is n_{ji} and the number of instances in class ω_i is n_i. Estimate the probability mass $p_{X|\omega_i}(x_j|\omega_i) = P(X = x_j|\omega_i)$ as n_{ji}/n_i for each discrete feature. Is this a valid estimate of the pmf?

(d) There is an issue with using the estimate you calculated in 4c. Explain why the laplace correction $(n_{ji}+1)/(n_i+l)$, where l is the number of levels X can assume,[1] solves the problem with the estimate given in 4c. Is this a valid estimate of the pmf?

(e) Estimate the minimum error rate decision rule for classifying tax evasion using Laplace correction.

5. **Programming Part: Breast Cancer Prognosis**

The goal of this assignment is to determine the prognosis of breast cancer patients using the features extrracted from digital images of Fine Needle Aspirates (FNA) of a breast mass. You will work with the Wisconsin Prognostic Breast Cancer data set, WPBC. There are 34 attributes in the data set: the first attribute is a patient ID, the second is an outcome variable that shows whether the cancer recurred after two years or not (N for Non-recurrent, R for Recurrent), the third variable is also an income

[1]For example, if $X \in \{apple, orange, pear, peach, blueberry\}$, then $d = 5$.

variable that shows the time to recurrence. The other 30 attributes are the features that you will work with to build a diagnosis tool for breast cancer.

Ten real-valued features are calculated for each nucleus in the digital image of the FNA of a breast mass.[2] They are:

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness (perimeter^2 / area - 1.0)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension "coastline approximation" - 1)

Ther mean, standard deviation, and the mean of three largest values for each image has been computed, to represent each image using 3×10 features.

Additionally, the diameter of the excised tumor in centimeters and the number of positive axillary lymph nodes are also given in the data set.

Important Note: Time to recurrence (third attribute) should *not* be used for classification, otherwise, you will be able to perfectly classify!

There are 198 instances in the data set, 151 of which are nonrecurrent, and 47 are recurrent.

(a) Download the WPBC data from: `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)`.

(b) Select the first 130 non-recurrent cases and the first 37 recurrent cases as your training set. Add record #197 in the data set to your training set as well. (10 pts)

(c) There are four instances in your training set that are missing the lymph node feature (denoted as ?). This is not a very severe issue, so replace the missing features with the median of the lymph node feature in *your training set*. (5 pts)

(d) Binary Classification Using Naïve Bayes' Classifiers

   i. Solve the problem using a Naïve Bayes' classifier. Use Gaussian class conditional distributions. Report the confusion matrix, ROC, precision, recall, F1 score, and AUC for both the train and test data sets. (10 pts)

[2]For more details see: `https://www.researchgate.net/publication/2512520_Nuclear_Feature_Extraction_For_Breast_Tumor_Diagnosis`.

ii. This data set is rather imbalanced. Balance your data set using SMOTE, by downsampling the common class in the training set to 90 instances and upsampling the uncommon class to 90 instances. Use $k = 5$ nearest neighbors in SMOTE. Remember not to change the balance of the test set. Report the confusion matrix, ROC, precision, recall, F1 score, and AUC for both the train and test data sets. Does SMOTE help? (10 pts)

(e) (Extra practice, will not be graded) Solve the regression problem of estimating time to recurrence (third attribute) using the next 32 attributes. You can use KNN regression. To do it in a principled way, select 20% of data points each class in your training set to choose the best $k \in 1, 2, \ldots, 20$, and the rest 80% as the *new training set*. Report your MSE on the test set using the k you found and the whole training set (not only the new training set!). For simplicity, use Euclidean Distance. Repeat this process when you apply SMOTE to your new training set to only upsample the rare class and make the data completely balanced. Does SMOTE help in reducing the MSE?