Pattern Recognition

Amirkabir University
of Technology
(Tehran Polytechnic)

Fall 2023

Assignment 2
Decision is the Key!

Homeworks Guidelines and Policies

- ***What you must hand in.*** It is expected that the students submit an assignment report (HW2_[student_id].pdf) as well as required source codes (.m or .py) into an archive file (HW2_[student_id].zip). Please combine all your reports just into a single .pdf file.
- ***Pay attention to problem types.*** Some problems are required to be solved *by hand* (shown by the ✎ icon), and some need to be implemented (shown by 🐍 the icon). Please do not use implementation tools when it is asked to solve the problem by hand, otherwise you will be penalized and lose some points.
- ***Don't bother typing!*** You are free to solve by-hand problems on a paper and include their pictures in your report. Here, cleanness and readability are of high importance. Images should also have appropriate quality.
- ***Reports are critical.*** Your work will be evaluated mostly by the quality of your report. Do not forget to explain your answers clearly, and provide enough discussions when needed.
- ***Appearance matters!*** In each homework, 5 points (out of a possible 100) belong to compactness, expressiveness, and neatness of your report and codes.
- ***MATLAB is also allowable.*** By default, we assume you implement your codes in Python. If you are using MATLAB, you have to use the equivalent functions when it is asked to use specific Python functions.
- ***Be neat and tidy!*** Your codes must be separated for each question, and for each part. For example, you have to create a separate .py file for part b. of question 3, which must be named 'p3b.py'. (or .ipynb)
- ***Use bonus points to improve your score.*** Problems with bonus points are marked by the ⭐ icon. These problems usually include uncovered related topics, or those that are only mentioned briefly in the class.
- ***Moodle access is essential.*** Make sure you have access to Moodle, because that is where all assignments as well as course announcements are posted. Homework submissions are only made through Moodle.

- ***Assignment Deadline.*** Please submit your work **before the end of November 25th**.
- ***Delay policy.*** During the semester, students are given only **7 free late days** which they can use them in their own ways. Afterwards, there will be a 20% penalty for every late day, and no more than four late days will be accepted.
- ***Collaboration policy.*** We encourage students to work together, share their findings, and utilize all the resources available. However you are not allowed to share codes/answers or use works from the past semesters. Violators will receive a zero for that particular problem.
- ***Any questions?*** If there is any question, please do not hesitate to contact us through the Telegram group chat or following email addresses: **m.ebadpour@aut.ac.ir** and **atiyeh.moghadam@aut.ac.ir**.

1. Be careful of counterfeit receipts! (15 Pts.)

Keywords: classification problem, Bayes decision rule, credit card fraud detection.

One of the main challenges online shops face these days is dealing with fake receipts, especially during discount events like Black Friday, when they are too swamped to verify the authenticity of each transaction. Developing a pattern recognition system that can accurately discern between genuine and counterfeit receipts would significantly ease the workload of online shop owners. In this problem, the Bayes decision rule proves to be an incredibly straightforward yet effective approach for classifying fraudulent and authentic transactions. Here, you are only allowed to use two features.

اپلیکیشن بله

انتقال پول
کارت‌ب‌کارت با موفقیت انجام شد
مبلغ:۱۲,۳۴۵,۶۷۸,۰۰۰
کارت مبدا:۶۰۲۸۸۹xxxxxx۶۷۳۴
کارت مقصد: ۶۰۳۷۹۹xxxxxx۶۱۴۷
صاحب کارت مقصد: درس پترن
شماره پیگیری۴۷۱۵۶۳۳۳۷۳۶۹
تاریخ انتقال:۱۰ آبان ۱۴۰۲ ۱۳:۵۷
#کارت_ب_کارت #واریز #موفق

دانلود از: https://bale.ai/dl

Figure 1: As you can see, fake receipt makers can create a variety of realistic-looking receipts and can even generate invoices on the carpet of your house!!

F1	F2	F3	F4	F5	label
-21.2	126.5	152.2	45.7	-205.1	genuine
64.6	125.1	315.9	-47	110.9	fraudulent
159.4	-177.8	36.9	-204.1	126.5	fraudulent
121.9	-252.2	-149.2	-177.1	61.3	fraudulent
-18.6	26.1	4.1	24.7	46.4	fraudulent
44.2	-99.8	-67.1	-65.7	-56.6	fraudulent
-45.8	54.3	-5.8	59.2	-77.1	genuine
-56.1	130.2	90.3	84.1	164.4	genuine
-138	179.7	7.5	181.3	-123	genuine
-89.9	42.8	-108.4	104.6	-63	genuine
-112.3	68.9	-112.1	133.6	227.8	genuine
-95.1	175.5	84.1	134.2	134.1	genuine

a. By visual inspection using 2D feature space, evaluate which two features are the most suited. Explain your reasons.

b. Design a classifier using the bayes rule by considering the two features you picked in the previous part. The data are assumed to have gaussian distributions with the same covariance matrix $\Sigma = I$. Find the general form of discriminant function.

c. Classify the following data points using the functions you obtained in the previous part.

F1	F2	F3	F4	F5
41.9	63.1	177.3	-33.7	35
-78.9	119.1	29.3	106.6	102.6
-66.4	-61.9	-222.9	60.4	-5.5
133	-171.9	-5.2	-174.6	-52.5

d. Suggest a cost function for this problem and explain why you believe your chosen cost function is appropriate.

e. Express some of the challenges this system encounters.

f. Suppose we have two normal distributions with the same covariances but different means: $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$. In terms of their prior probabilities P(ω1) and P(ω2), state the condition that the Bayes decision boundary not pass between the two means.

2. How can I become over overproductive? (20 Pts.)

Keywords: parameter estimation, maximum likelihood estimation, Poisson distribution, unbiased estimator, estimator variance, consistent estimator, posterior probability, prior probability, probability of error

In today's fast-paced world, there is a prevalent obsession with productivity and effective time management. Many of us strive to maximize our time by staying up late, rising early, and establishing routines and habits, sometimes attempting to fit more than 24 hours into a day! However, with the increasing shift towards online work, managing time has become a greater challenge. Balancing ongoing projects after regular working hours, attending online meetings, and handling a flood of notifications and messages can make it feel like we have less control over our time.

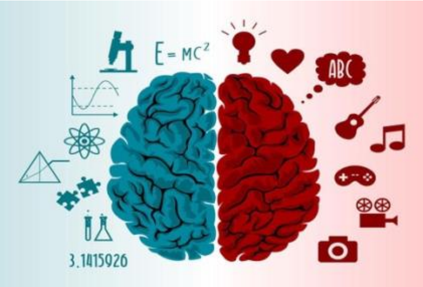

Figure 2: The left hemisphere is typically associated with analytical thinking, logic, and language. It is also responsible for planning and organizing tasks. When the left hemisphere is working well, we are able to focus on our goals, break down complex tasks into manageable steps, and see them through to completion.

Imagine wanting to allocate a specific time for responding to Telegram messages and emails, in order to free up time for other activities. You know that responding to each message takes approximately 3 minutes, but you have no way of predicting how many messages you might receive within a 6-hour window. This scenario can be modelled using a Poisson distribution to estimate the number of messages received over a specific time period.

Suppose that $X_1, X_2, \dots, X_n, n > 5$ forms an iid (random) sample from a Poisson distribution with parameter λ:

$$f(x|\beta) = \frac{\lambda^x e^{-\lambda}}{x!}, \qquad x = 0,1,2, \dots .$$

Consider the following estimators of λ:

$$\widehat{\lambda_1} = X_n; \qquad \widehat{\lambda_2} = \frac{1}{5}\sum_{i=1}^{5} X_i\,; \qquad \widehat{\lambda_3} = \frac{1}{n}\sum_{i=1}^{n-1} X_i.$$

a. Which, if any, of these three estimators are unbiased?
b. Find the variance of these three estimators. Which, if any, of these three estimators are consistent? Modify the biased but consistent estimators to make them unbiased.
c. Find the maximum likelihood estimator of λ, λ_{MLE}. Show that λ_{MLE} is an unbiased estimator of λ.
d. The number of messages received within a 6-hour window in 3 days is observed as follows:

0,3,4,7,3,2,1,1,2,3,4,0,2,5,2,1

Calculate the maximum likelihood estimator of λ and $\log(\lambda)$.

e. Let $p(x|\omega_i) \sim N\ (\mu_i\ , \sigma^2)$ for a two-category one-dimensional problem with $P(\omega_1) = P\ (\omega_2) = \frac{1}{2}$. Show that the minimum probability of error is given by

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

where $a = \frac{|\mu_2 - \mu_1|}{2\sigma}$.

f. Use the inequality

$$P_e = \frac{1}{\sqrt{2\pi}} \int_a^{\infty} e^{-\frac{u^2}{2}}\, du \ \le \ \frac{1}{\sqrt{2\pi} a} e^{-\frac{a^2}{2}}$$

to show that P_e goes to zero as $\frac{|\mu_2 - \mu_1|}{\sigma}$ goes to infinity. Express the meaning of this result in words.

3. Is It Worth Using Dokkan? (20 Pts.)

Keywords: *parameter estimation, maximum likelihood estimation, geometric distribution, beta distribution, posterior probability, prior probability*

When people decide to launch an online shop, a primary question arises regarding which platform to choose. There are various options available, such as Instagram, Basalam, or having your own website. Now, a new platform called Dokkan offers a plan where they not only sell your products but also run online advertisements on popular Persian customer-centric sites. However, it's worth noting that advertising comes with an additional cost. For a newly launched business, investing in online advertisements may entail some risk.

Figure 3 parameter estimation can be used to estimate the effectiveness of different ad formats, targeting strategies, and bidding algorithms. This information can then be used to improve the performance of future campaigns.

Dokkan provides you with data regarding the number of ad impressions before a user clicks on an advertisement. Estimating the parameters of distributions is useful for modelling and comprehending customer behaviours. The list below comprises 25 numbers, and it is known that this data follows a geometric distribution with an unknown parameter p:

3, 4, 3, 3, 2, 3, 2, 7, 10, 2, 5, 3, 3, 8, 1, 1, 1, 6, 5, 6, 11, 5, 2, 5, 1

a. plot the bar graph which indicates the values in the list and their number of occurrences. Normalize the plot and plot the empirical probability mass function.

b. plot the PMF of a geometric distribution with the following parameter p:

p = 0.2 p = 0.05 p = 0.7

c. which of the PMFs in part (b) is more likely to have generated the given list of 25 numbers? plot the likelihood function of the given numbers as a function of the parameter p. also plot the log likelihood function. determine the value of p that maximizes both likelihood and log likelihood.

Next, you are provided with a dataset containing 5000 random samples generated from a beta distribution. This distribution is used for modelling the proportion of website visitors who make a purchase. The beta distribution is defined as follows:

$$f(x;a,b) = \frac{\Gamma(a+b)}{\Gamma(\mathrm{a})\Gamma(\mathrm{b})}\, x^{a-1}(1-x)^{b-1}$$

d. Write down the likelihood and log-likelihood functions.

e. Estimate a and b by approximately maximizing the log-likelihood function. (range: [0, 15]). Calculate the exact values of a and b using an arbitrary optimization approach, either from a built-in function or an external library.

f. Plot and compare the estimated functions from part (f) with the histogram of the given samples

4. To Keep or Let Go: A Customer Churn Problem (15 Pts.)

Keywords: *Quadratic discriminant analysis, confusion matrix, precision, recall, true positive rate, false negative rate, classification evaluation, receiver operating characteristic, ROC, area under curve, AUC*

In the competition among various businesses operating in a specific domain, identifying dissatisfied customers and those on the verge of discontinuing their services is of great importance. For various reasons, a customer may choose to cancel their use of the provided service at any given time. These reasons may include dissatisfaction with the quality of the service, price increases, or finding a similar service with better conditions from another company or provider. By utilizing pattern recognition and data analysis, we can identify these customers and take appropriate measures to retain them, ensuring they won't leave our service. This pattern recognition problem is referred to as customer churn prediction.

Figure 4 Customer churn, also known as customer attrition, is the rate at which customers stop doing business with a company. It's a critical metric for businesses to track, as high churn rates can lead to significant financial losses and hinder growth.

Suppose we have data regarding customers likely to stay in X1train, and customers likely to leave in X2train. In the first step, we aim to use Quadratic Discriminant Analysis for classification.

$$\text{X1train} = \begin{bmatrix} 0 & 0 & 1 & 2 & 3 & 2 & 2 & 2 & 1 & 4 \\ 1 & 3 & -1 & 0 & 2 & 3 & 1 & 2 & 0 & 0 \end{bmatrix}$$

$$\text{X2train} = \begin{bmatrix} -3 & -3 & -2 & -1 & -2 & -2 & 1 & 0 & 1 & -1 \\ 0 & 2 & -2 & 0 & 1 & 2 & 2 & 2 & 1 & 1 \end{bmatrix}$$

a. Calculate the prior distribution for each class, along with the mean vector and covariance matrix. [you can use online calculators for calculating the covariance matrix.]

b. Define the likelihood of each class as a multivariate Gaussian distribution using the mean vector and covariance matrix calculated in the previous step. Write down the formula for the posterior probability of each class. What do you anticipate the decision boundary to look like? Could you provide the formula for the decision boundary?

c. Define 'p' as p = $\frac{posterior\ of\ class\ P}{posterior\ of\ class\ P + posterior\ of\ class\ N}$ Then, classify samples as 'P' if p[i] > threshold, else 'N'. Calculate the value of 'p' for the test samples below.

$$\text{Xtest} = \begin{bmatrix} 8 & 3 & 1 & 4 & 0 & 4 & 0 & -2 & 0 & -3 \\ 5 & 0 & -2 & 1 & 2 & 4 & 0 & -3 & -2 & 1 \end{bmatrix}$$

Given the significance of this problem for businesses, service providers are keen on identifying the best model with the highest number of accurate predictions. Suppose two other classification models were employed for customer classification ('P' for customers likely to stay, and 'N' for customers likely to cancel their use of the service). The actual observed outcomes and predicted probabilities for the test samples in the previous part are provided in the following table.

Test data	Actual outcome	Model 1 prediction	Model 2 prediction
(8, 5)	P	0.7	0.75
(3, 0)	P	0.8	0.8
(1, -2)	P	0.65	0.65
(4, 1)	P	0.9	0.85
(0, 2)	P	0.45	0.3
(4, 4)	N	0.5	0.45
(0, 0)	N	0.55	0.55
(-2, -3)	N	0.35	0.35
(0, -2)	N	0.4	0.4
(-3, 1)	N	0.6	0.25

d. For each model including your own, create a confusion matrix with a threshold of 0.5 and then calculate FPR (false positive rate), Accuracy, Precision, Recall, and F1 score.Which of the metrics you calculated is more important in this problem? If we intend to use the f_beta measure, what would be a suitable value for beta?

e. Draw two ROC curves: one for Model 1 and another for Model 2, preferably in one figure using different colours (e.g., Red for model 1 and Blue for model 2). Compare the AUC (area under the curve) for each ROC curve and conclude whether one model is superior to the other.

f. Suppose you developed a new model and wanted to show that yours is dominantly superior to the other two models. What are the desired properties of your model's ROC curve? Provide a sample set of probabilities that your model may generate.

g. Can we improve any of the previous scores (without a negative effect on any of the other scores) by changing the threshold? If yes, which threshold value would you choose and why? If not, explain why not.

5. Think Complex as graph! (25 +4 Pts.)

Keywords: Non-Stationary data distributions, Graph Structures, *Node Classification, Minimum Distance Classifier, Neighbourhood Extraction, Breadth-First Search, Exponential Weighting.*

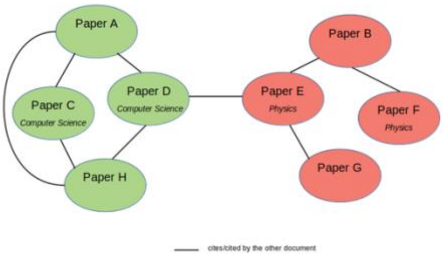

Figure 5: Sample schema of scientific publication graph.

Node classification is a fundamental task in graph machine learning, aiming to predict the category or label of each node in a graph. The **Cora dataset**, a citation network of machine learning papers, is widely used to benchmark node classification algorithms. Accurately classifying nodes in the Cora dataset is crucial for various applications, including topic classification of academic papers, recommendation of relevant research articles, and identification of influential publications. Effective node classification techniques can help researchers better understand the relationships between scientific papers, identify emerging trends in research, and make informed decisions about their own work.

The Cora dataset consists of **2708 scientific publications (nodes)** classified into one of **seven classes**. The citation network consists of **5429 links**. Each publication in the dataset can be described by a **0/1-valued word vector** indicating the absence/presence of the corresponding word from the dictionary. The dictionary consists of **1433 unique words**.

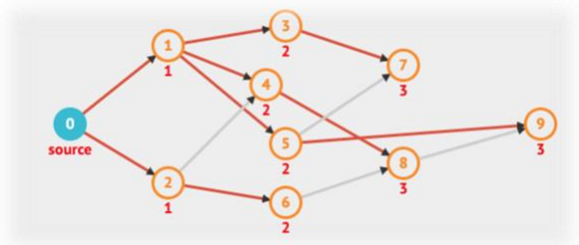

Figure 6: Sample of BFS algorithm. Level of each nieghbor is indicated.

In this problem, we want to implement a **node classification** algorithm using the **minimum distance classifier**. Evaluate the performance of the algorithm using various **neighborhood** levels and compare the results. Additionally, apply the **exponential weighting** mechanism to the node classification task and discuss its effectiveness.

a) Explore the Cora dataset (attached) and convert it into a binary representation where each row represents a node and each column represents the presence or absence of a specific word from a predefined vocabulary.

b) For each node in the dataset, perform classification using its own feature vector and the feature vectors of its direct neighbors. For each unique class among its neighbors, calculate the prototype based on the corresponding feature vectors. Assign the label of the nearest prototype to the test node using the Euclidean distance. Report the Top-1 and Top-5 accuracy as well as confusion matrix for the classification task.

Breadth-First Search (BFS) is an algorithm for traversing graph data structures. It starts at the source node and explores the neighbor nodes first, before moving to the next level neighbors. In other words, it visits all nodes at a depth *d* in the graph before visiting any nodes at depth *d+1*. This can be implemented using a **simple queue data structure**. The nodes are added to the queue in order they are discovered, and the queue is popped to retrieve the next node to visit.

c) Repeat part b. but extract the feature vectors of neighbors up to three levels using a breadth-first search (BFS) algorithm. Compare the results multi-level neighbor classification with the one-level neighbor classification.

Note: When extracting neighbors using BFS, it is possible to encounter a loop, where a node is its own neighbor. In this case, we should not use the ground truth label of the source node in the calculation of the prototype, as this would lead to a biased result.

d) Plot a chart that x-axis is number of level of neighbors that used and y-axis is the 1-Top accuracy. Based on obtained chart, discuses and select a level based on experience. (For each level you should repeat classification)

e) Describe the exponential weighting mechanism and its role in adapting to **non-stationary data distributions** (what is the non-stationary data distributions?). Apply exponential weighting to the node classification task, considering different neighborhood levels (e.g., 3-level, 5-level, 7-level). Explain the strategy for configuring the parameters of the exponential weighting mechanism.

f) Discuss why extracting static prototypes for each class from the entire dataset may not be suitable for real-world applications. Explain how dynamic prototypes, which adapt to local neighborhood information, can be more effective in real-world scenarios.

Good Luck!

Mohsen Ebadpour, Atiyeh *Moghadam, Romina Zakerian*