Problem Set 1 — PMR3508 (2021)

Prof. Fabio Cozman

1) What is:

i) Supervised learning? Unsupervised learning? Semi-supervised learning?

ii) Feature? Label?

iii) Missing data? Missing data at random?

iv) Training dataset? Testing dataset? Validation dataset?

v) Data preparation? What does it worry about?

vi) Classification? Regression?

vii) Classifier?

viii) Error rate? Empirical error rate?

ix) Overfitting? Underfitting?

x) Cross-validation? Leave-one out cross-validation? Stratified cross-validation?

xi) Bayes classifier? Bayes error?

xii) Plug-in classifier?

xiii) Nearest-neighbor classifier?

xiv) Expected quadratic error? Mean squared error?

xv) Bias? Variance? Bias-variance trade-off?

xvi) Maximum likelihood estimation?

xvii) Confusion matrix? True positive? False positive? False negative? True negative?

xviii) Precision? Recall? Sensitivity? Specificity?

xix) True positive rate? False positive rate? False negative rate? True negative rate?

xx) ROC curve? Area under the curve? What is the best classifier with respect to the ROC curve? What is the meaning of the diagonal from $(0, 0)$ to $(1, 1)$?

2) Solve Problem 7, Chapter 2 in *An Introduction to Statistical Learning*.

3) The probability that team A loses is $1/4$; the probability that the coach is replaced given that team A loses is $9/10$; and the probability that the coach is replaced given that team A wins is $3/10$. What is the probability that team A loses given that the coach is replaced?

4) Consider a random variable X with integer values $-5, -4, -3, \ldots, 3, 4, 5$, with uniform distribution. Consider another random variable $Y = X^2$. What is the expected value of Y? What is the probability $\mathbb{P}(X > 0|Y > \pi)$?

5) The table below conveys the joint probability mass function for discrete random variables X and Y.

	y_1	y_2	y_3	y_4
x_1	$0,125$	a	$0,10$	$0,125$
x_2	$0,05$	$0,06$	b	$0,05$
x_3	c	$0,09$	$0,06$	$0,075$

If X and Y are independent, what are the values of a e b e c?

6) Consider two random variables X and Y; X has values $\{1,2,3\}$ and Y has values $\{-1,0,1\}$. Suppose that the joint distribution of X and Y is specified by $\mathbb{P}(X=x, Y=y) = k(x+y+n)$ for $x \in \{1,2,3\}$ and $y \in \{-1,0,1\}$, where n is the last digit of your USP number and k is a constant.

a) Obtain the constant k.

b) Obtain $\mathbb{P}(X=2)$ and $\mathbb{P}(Y=0|X=2)$.

c) Obtain the decision of the Bayes classifier for class variable Y when X is the only feature and the observation is equal to 2.

7) Study Example 1.3, Chapter 1 in *Bayesian Reasoning and Machine Learning*.

8) Consider two binary variables X (with values 10 and 20) and Y (with values 0 and 1). Suppose we know that $\mathbb{P}(Y=0) = 1/4$, $\mathbb{P}(X=10|Y=0) = 1/5$ and $\mathbb{P}(X=10|Y=1) = 7/8$.

a) What is the value of $\mathbb{P}(Y=1|X=10)$?

b) What is the value of $\mathbb{P}(X=10)$?

c) Suppose you observe X and you have to classify this observation by selecting a value of Y. So your classifier is the function $g(X)$, with possible values 0 and 1. Suppose you want to use the *best* possible classifier with respect to error rate. If you observe $\{X=10\}$, what is the value of $g(10)$?

9) Suppose we have a class variable Y with values 0 and 1, and a feature X with integer values. Suppose:

$$\mathbb{P}(X=x|Y=0) = \binom{5}{x} (1/3)^x (2/3)^{5-x}, \qquad \text{for } x \in \{0,1,\ldots,5\}$$

and

$$\mathbb{P}(X=x|Y=1) = \binom{5}{x-2} (3/4)^{x-2} (1/4)^{5-(x-2)}, \qquad \text{for } x \in \{2,6,\ldots,7\};$$

also, $\mathbb{P}(Y=1) = 1/2$. What is the Bayes classifier? What is the Bayes error?

10) Suppose we have a class variable Y with values 0 and 1, and two features X_1 and X_2 such that:

$$\mathbb{P}(X_1=a, X_2=b|Y=y) = \begin{cases} \alpha(a+b)y + \beta\frac{1-y}{ab} & \text{for } a \in \{1,2,3\}, b \in \{1,2,3\}, \\ 0 & \text{otherwise,} \end{cases}$$

where α and β are constants. Additionally, $\mathbb{P}(Y=0) = 2/3$. Suppose that we collect three observations:

X_1	X_2	Y
1	1	0
1	2	0
3	3	1

a) What is the value of α and β?

b) What is the Bayes classifier? What is the Bayes error?

c) Suppose we build a 1NN classifier with Euclidean distance (where ties are resolved randomly with identical probabilities for labels 0 and 1). What is the error rate of this classifier?

11) Suppose we have a classifier with class variable Y consisting of values 0 and 1; we are interested in determining when an object gets label 1. Suppose also that we made a few experiments and we determined that

- we had 100 cases where the object was 0 and was labeled 0;
- we had 10 cases where the object was 0 and was labeled 1;
- we had 5 cases where the object was 1 and was labeled 0;
- we had 150 cases where the object was 1 and was labeled 1.

a) What is the confusion matrix for this problem?

b) Obtain precision/recall/F1 for this classifier.

12) Consider a classifier with binary output Y; suppose the following confusion matrix is computed using a test dataset:

	Actual: 1	Actual: 0
Classified: 1	12	10
Classified: 0	4	250

a) What is the accuracy of this classifier?

b) What is the precision and the recall?

c) What is the F_1 score?

d) Consider an alternative classifier that would output $\{Y = 0\}$ for every input, and another classifier that would output $\{Y = 1\}$ for every input. Suppose we use the same testing data. What is the accuracy of these classifiers? What is their precision and the recall? What is their F_1 score?

13) Consider a classifier with a binary feature X and a binary class variable Y. Suppose we know that $\mathbb{P}(Y = 0) = \alpha$. Suppose also that

$$\mathbb{P}(X = 1|Y = 0) = 2\theta, \qquad \mathbb{P}(X = 1|Y = 1) = \beta.$$

Suppose we have the following training dataset:

x	0	0	1	0	1	0	0	1	0	0
y	1	0	1	0	0	0	0	1	0	1

a) What is the likelihood function for parameter θ as a function of α and β?

b) What is the maximum likelihood estimate of θ as a function of α and β?

We want to build a plug-in classifier whose output $\hat{Y}$ must be selected so as to minimize expected cost $\mathbb{E}\left[c(Y,\hat{Y})\right]$, where the function $c(Y,\hat{Y})$ is given by

	Actual $Y=1$	Actual $Y=0$
Classified $\hat{Y}=1$	0	5
Classified $\hat{Y}=0$	10	1

a) What is this classifier for $\alpha = 2/3$ and $\beta = 1/4$?

b) What is the accuracy of this classifier? What is its precision, its recall, and its F_1 score?

14) A classifier was tested on a testing dataset consisting of 1000 observations, producing 250 true positives, precision 0.9 and recall 0.8. Determine a confidence interval for the classifier error rate, with confidence 0.95.

15) Suppose we have a class variable Y with values 0 and 1, such that $\mathbb{P}(Y=1) = 1/2$. We also have a feature X that has Gaussian distribution conditional on Y: if $\{Y = y\}$, then X has Gaussian distribution with expected value y and variance 1. Suppose a classifier yields, for an observed x, $\hat{Y} = 1$ if $\mathbb{P}(Y=1|x)/\mathbb{P}(Y=0|x) > \alpha$ and $\hat{Y} = 0$ otherwise (where α is a design parameter larger than zero).

Draw the ROC curve for this classifier (plot at least five points).

16) Two classifiers were built using maximum likelihood estimation with two different prior hyperparameters. Classifiers C_1 and C_2 were tested with 10-fold cross-validation on a testing dataset of total size equal to 2000. The resulting empirical error rates for the folds are:

C_1	0.9	0.89	0.87	0.9	0.88	0.9	0.91	0.89	0.88	0.89
C_2	0.89	0.87	0.86	0.89	0.92	0.88	0.87	0.9	0.9	0.87

Apparently classifier C_1 is better than C_2. Using a Wald test, decide whether the null hypothesis that they are equal can indeed be rejected at significance level 0.05. Can it be rejected at significance 0.1?

Remember: the Wald statistic with respect to accuracies a_i and b_i is

$$W = \frac{\sum_{i=1}^{N}(a_i - b_i)/N}{\sqrt{\frac{1}{N(N-1)}\sum_{i=1}^{N}\left(a_i - b_i - \frac{\sum_{j=1}^{N}(a_j-b_j)}{N}\right)^2}}.$$