Problem Set 1 — PMR3508

Prof. Fabio Cozman — 2^o semester 2019

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 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) Naive Bayes classifier?

xvii) Maximum likelihood estimation? Bayesian estimation?

xviii) Expectation-Maximization (EM) algorithm?

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

xx) Precision? Recall? Sensitivity? Specificity?

xxi) True positive rate? False positive rate? False negative rate? True negative rate?

xxii) 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 $(1, 1)$?

xxiii) An artificial neural network?

xxiv) A perceptron?

xxv) The function encoded by a perceptron?

xxvi) The learning rule for the perceptron?

xxvii) The disadvantages of the perceptron as classifier?

xxviii) A multilayer perceptron?

xxix) A hidden layer?

xxx) The recall of a multilayer perceptron?

xxxi) The sigmoid logistic function?

xxxii) The backpropagation algorithm?

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

3) 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 $\mathbb{P}(X > 0 | Y > \pi)$?

4) 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?

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

6) 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 $g(10)$?

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

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

and

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

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

8) 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?

9) Consider a Naive Bayes classifier, where attributes X_1 and X_2 have values 0 and 1. The label variable Y has values standing for spam and text.

This classifier is used to classify text in a particular company. There is a large dataset with classified documents in this company. The classifier was trained with the dataset, thus producing estimates $\mathbb{P}(Y = \text{spam}) = 0.1$, $\mathbb{P}(X_i = 0|Y = \text{spam}) = 0.3$ and $\mathbb{P}(X_i = 0|Y = \text{text}) = 0.8$ (both for X_1 and X_2).

a) If we observe $\{X_1 = 0\}$, what is $\mathbb{P}(Y = 1|X_1 = 0)$?

b) Suppose this Naive Bayes classifier is used as a plug-in classifier; what is then the label produced by the classifier given observed attributes $\{X_1 = 0, X_2 = 0\}$?

c) The company hires you as a consultant, and asks whether this plug-in classifier is guaranteed to be the best possible classifier for them. You say no. They ask why, and they ask you what would be the best possible classifier. What is the answer to them?

10) Solve Problem 10.1, Chapter 10 in *Bayesian Reasoning and Machine Learning.*

11) Solve Problem 10.2, Chapter 10 in *Bayesian Reasoning and Machine Learning.*

12) Solve Problem 10.3, Chapter 10 in *Bayesian Reasoning and Machine Learning.*

13) Solve Problem 10.4, Chapter 10 in *Bayesian Reasoning and Machine Learning.*

14) A doctor wants to learn the value of the probability that a person contracts zika in a particular city. This probability is denoted by θ. The doctor holds a prior distribution over θ that is a Beta distribution with parameters $\alpha = \beta = 2$. The doctor randomly observes people in the city, noting that out of 1000 people, only 18 have zika.

a) What is the likelihood as a function of θ?

b) What is the posterior distribution of θ?

c) What is the conditional expectation a posteriori estimate of θ?

15) Consider a Naive Bayes classifier, with two attributes X_1 and X_2, and a class variable Y. Suppose all random variables have values 0 and 1. The following data was collected.

X_1	X_2	Y
0	1	0
0	0	0
0	1	1
1	1	1
1	0	1
0	1	1

a) Write down the parameters of the Naive Bayes that need to be estimated from data.

b) What are the maximum likelihood estimates for the parameters?

c) Consider a Bayesian analysis where all parameters are independent. Suppose the prior distribution for each parameter is Beta, all with parameters $\alpha = \beta = 2$. What is the posterior distribution for each parameter?

d) What are the estimates produced by conditional expectation a posteriori for the parameters?

16) 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

What is the accuracy of this classifier? What is the precision and the recall? What is the F_1 score?

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?

17) 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

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

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 $\mathbb{E}\Big[c(Y, \hat{Y})\Big]$, 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

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

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

18) 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.

19) 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).

20) Two Naive Bayes classifiers were built using Bayesian 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

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

21) Consider a perceptron with three inputs, X_1, X_2 and X_3, and a continuous output Y that is produced by a logistic activation function $f(x_1, x_2, x_3)$ of the weighted sum of inputs W:

$$Y = \frac{1}{1 + \exp(-(1/3)W)},$$

where W is computed with weights 1, 2, and 4 for inputs X_1, X_2 and X_3 respectively.

What is the output Y for inputs $(x_1, x_2, x_3) = (-3, 10, -2)$?

What is the output Y if the logistic activation function is replaced by a ReLU activation function?

Table of values for $\int_0^{z_0}(1/\sqrt{2\pi})\exp(-z^2/2)dz$.

Distribuição Normal P(0≤Z<z0)										
z0	**0,00**	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990