Aprendizagem 2021

Lab 2: Probability theory and Bayesian learning

Practical exercises

I. Evaluation (*cont.*)

1. Consider 7 observations with the following annotations $\boldsymbol{y} = [0\ 1\ 0\ 1\ 0\ 0\ 1]$, and the probabilistic outcome of a classifier under a one-leave-out scheme $p(\boldsymbol{y} = 1|\boldsymbol{x}) = [0.55\ 0.8\ 0.4\ 0.6\ 0.6\ 0.3\ 0.9]$.

a) Draw the ROC curve using an 11-point interpolation

b) Compute the AUC

c) Would you change the default 0.5 probability threshold for this classifier?

II. Probability theory

2. Consider the following registry where an experiment is repeated six times and four events (A, B, C and D) are detected. Considering frequentist estimates, compute:

$p(A)$
$p(A, B)$
$p(B|A)$
$p(A, B, C)$
$p(A|B, C)$
$p(A, B, C, D)$
$p(D|A, B, C)$

	A	B	C	D
x_1	1	1	0	0
x_2	1	1	1	0
x_3	0	0	0	1
x_4	0	0	0	1
x_5	0	0	0	0
x_6	0	0	0	0

3. Considering the following two-dimensional measurements {(-2,2),(-1,3),(0,1),(-2,1)}.

a) What are the maximum likelihood parameters of a multivariate Gaussian distribution for this set of points?

b) What is the shape of the Gaussian? Draw it approximately using a contour map.

III. Bayesian learning

4. Consider the following dataset where:

- 0: False and 1: True
- y1: Fast processing
- y2: Decent Battery
- y3: Good Camera
- y4: Good Look and Feel
- y5: Easiness of Use
- class: iPhone

	y_1	y_2	y_3	y_4	y_5	class
x_1	1	1	0	1	0	1
x_2	1	1	1	0	0	0
x_3	0	1	1	1	0	0
x_4	0	0	0	1	1	0
x_5	1	0	1	1	1	1
x_6	0	0	1	0	0	1
x_7	0	0	0	0	1	1

And the query vector $\boldsymbol{x}_{\text{new}} = [1\ 1\ 1\ 1\ 1]^T$

a) Using Bayes' rule, without making any assumptions, compute the posterior probabilities for the query vector. How is it classified?

b) What is the problem of working without assumptions?

c) Compute the class for the same query vector under the naive Bayes assumption.

d) Consider the presence of missings. Under the same naive Bayes assumption, how do you classify $\boldsymbol{x}_{\text{new}} = [1\ ?\ 1\ ?\ 1]^T$

5. Consider the following dataset

	weight (kg)	height (cm)	NBA player
x_1	170	160	0
x_2	80	220	1
x_3	90	200	1
x_4	60	160	0
x_5	50	150	0
x_6	70	190	1

And the query vector $\boldsymbol{x}_{\text{new}} = [100\ 225]^T$

a) Compute the most probable class for the query vector assuming that the likelihoods are 2-dimensional Gaussians

b) Compute the most probable class for the query vector, under the Naive Bayes assumption, using 1-dimensional Gaussians to model the likelihoods

6. Assuming training examples with *m* Boolean features.

a) How many parameters do you have to estimate considering features are Boolean and:

   i. no assumptions about how the data is distributed

   ii. naive Bayes assumption

b) How many parameters do you have to estimate considering features are numeric and:

   iii. multivariate Gaussian assumption

   iv. naive Bayes with Gaussian assumption

Programming quests

1. Reuse the **sklearn** code from last practical class where we applied a kNN to the iris dataset.
   a) apply the naïve Bayes classifier with default parameters.
   b) compare the accuracy of both classifiers using a 10-fold cross-validation.

2. Consider the accuracy estimates collected under a 5-fold CV for two predictive models M1 and M2, acc_{M1}=(0.7,0.5,0.55,0.55,0.6) and acc_{M2}=(0.75,0.6,0.6,0.65,0.55).
   Using **scipy**, assess whether the differences in predictive accuracy are statistically significant.
   Resource: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_rel.html