Aprendizagem 2022

Lab 3: Bayesian learning

Practical exercises

I. Probability theory

1. 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:

	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

$p(A) = \frac{2}{6}$

$p(A, B) = \frac{2}{6}$

$p(B|A) = 1$

$p(A, B, C) = \frac{1}{6}$

$p(A|B, C) = 1$

$p(A, B, C, D) = 0$

$p(D|A, B, C) = 0$

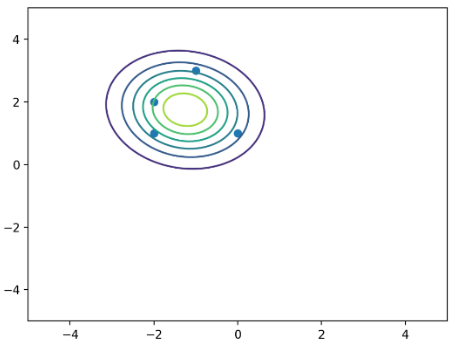
2. Considering the following two-dimensional measurements {(-2,2),(-1,3),(0,1),(-2,1)}.

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

$$N(\mathbf{x}|\mu, \Sigma), \quad \mu = \begin{bmatrix} -1.25 \\ 1.75 \end{bmatrix}, \quad \Sigma = \begin{pmatrix} 0.92 & -0.083 \\ -0.083 & 0.92 \end{pmatrix}, \quad det(\Sigma) = 0.83, \quad \Sigma^{-1} = \begin{pmatrix} 1.1 & 0.1 \\ 0.1 & 1.1 \end{pmatrix}$$

b) What is the shape of the Gaussian?

Draw it approximately using a contour map.



II. Bayesian learning

3. 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?

$$p(C=0)=\frac{3}{7}, \qquad p(C=1)=\frac{4}{7}$$

$$p\,(C=0 \mid y1=1, y2=1, y3=1, y4=1, y5=1)=\frac{p\,(C=0)p(y1=1, y2=1, y3=1, y4=1, y5=1|C=0)}{p\,(y1=1, y2=1, y3=1, y4=1, y5=1)}$$

$$p\,(C=1 \mid y1=1, y2=1, y3=1, y4=1, y5=1)=\frac{p\,(C=1)p(y1=1, y2=1, y3=1, y4=1, y5=1|C=1)}{p\,(y1=1, y2=1, y3=1, y4=1, y5=1)}$$

According to our estimated likelihoods, the denominators are equal to zero. Posteriors are not defined and, thus, we cannot classify the input. A small training sample is not enough to decide under a classic Bayes rule.

b) What is the problem of working without assumptions?

Insufficient data to construct a meaningful joint distribution, e.g. applicable for datasets with high dimensionality or low size (small sample).

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

$$p\,(C=0 \mid y1=1, y2=1, y3=1, y4=1, y5=1)=\frac{0.0141}{p(y1=1, y2=1, y3=1, y4=1, y5=1)}$$

$$p\,(C=1 \mid y1=1, y2=1, y3=1, y4=1, y5=1)=\frac{0.0090}{p(y1=1, y2=1, y3=1, y4=1, y5=1)}$$

Label $C=0$ (not an iPhone).

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

$$p\,(C=0 \mid y1=1, y2=?, y3=1, y4=?, y5=1)=\frac{0.03175}{p(y1=1, y3=1, y5=1)}$$

$$p\,(C=1 \mid y1=1, y2=?, y3=1, y4=?, y5=1)=\frac{0.0714}{p(y1=1, y3=1, y5=1)}$$

Label $C=1$.

4. 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.

$$p(C=0)=\frac{1}{2}, \qquad p(C=1)=\frac{1}{2},$$

$$p(y1, y2 \mid C=0) \qquad p(y1, y2 \mid C=1)$$

$$\mu \quad \begin{bmatrix} 93.(3) \\ 156.(6) \end{bmatrix} \qquad \begin{bmatrix} 80 \\ 203.(3) \end{bmatrix}$$

$$\Sigma \quad \begin{bmatrix} 4433.(3) & 216.(6) \\ 216.(6) & 33.(3) \end{bmatrix} \qquad \begin{bmatrix} 100 & 50 \\ 50 & 233.(3) \end{bmatrix}$$

$$p\,(C=0 \mid y1=100, y2=225) = \frac{p\,(C=0)p(\,y1=100, y2=225 \mid C=0)}{p(\,y1=100, y2=225)}$$

$$\frac{\frac{1}{2}N\left(\begin{bmatrix}100\\225\end{bmatrix} \middle| \mu=\begin{bmatrix}93.(3)\\156.(6)\end{bmatrix}, \Sigma=\begin{bmatrix}4433.(3) & 216.(6)\\216.(6) & 33.(3)\end{bmatrix}\right)}{p(\,y1=100, y2=225)} = \frac{1.74\times10^{-48}}{p(\,y1=100, y2=225)}$$

$$p\,(C=1 \mid y1=100, y2=225) = \frac{p\,(C=1)p(\,y1=100, y2=225 \mid C=1)}{p(\,y1=100, y2=225)}$$

$$\frac{\frac{1}{2}N\left(\begin{bmatrix}100\\225\end{bmatrix} \middle| \mu=\begin{bmatrix}80\\203.(3)\end{bmatrix}, \Sigma=\begin{bmatrix}100 & 50\\50 & 233.(3)\end{bmatrix}\right)}{p(\,y1=100, y2=225)} = \frac{5.38\times10^{-5}}{p(\,y1=100, y2=225)}$$

Classified as an NBA player.

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

	$p(y1 \mid C=0)$	$p(y1 \mid C=1)$	$p(y1 \mid C=0)$	$p(y1 \mid C=1)$
μ	$93.(3)$	80	$156.(6)$	$203.(3)$
σ	66.58	10	5.77	15.275

$$p\,(C=0 \mid y1=100, y2=225) = \frac{p\,(C=0)p(\,y1=100 \mid C=0)p(y2=225 \mid C=0)}{p(\,y1=100, y2=225)}$$

$$\frac{\frac{1}{2}N(100 \mid \mu=93.(3), \sigma=66.58)N(225 \mid \mu=156.(6), \sigma=5.77)}{p(\,y1=100, y2=225)} = \frac{7.854\times10^{-35}}{p(\,y1=100, y2=225)}$$

$$p\,(C=1 \mid y1=100, y2=225) = \frac{p\,(C=1)p(\,y1=100 \mid C=1)p(y2=225 \mid C=1)}{p(\,y1=100, y2=225)}$$

$$\frac{\frac{1}{2}N(100 \mid \mu=80, \sigma=10)N(225 \mid \mu=203.(3), \sigma=15.275)}{p(\,y1=100, y2=225)} = \frac{2.578\times10^{-5}}{p(\,y1=100, y2=225)}$$

Classified as an NBA player.

5. Assuming training examples with *m* features and a binary class.

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

One parameter for the prior $p(z=0)=1-p(z=1)$.

Considering the classic Bayesian model: we need $(2^m-1)\times 2$ parameters to estimate $p(y_1=v_1,\ldots,y_m=v_m \mid z=c)$, hence $2^m\times 2-1$.

Considering the naïve Bayes: we need to estimate $p(y_i \mid z=c)$. Since there are 2 classes and m features, we have $2\times m\times 1=2m$ parameters for the likelihoods. The total number of parameters is $1+2m$.

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

   iii. multivariate Gaussian assumption

   iv. naive Bayes with Gaussian assumption

Similarly, one parameter for the prior, $p(z=0) = 1 - p(z=1)$.

A multivariate Gaussian to estimate the likelihood $p(\mathbf{x} \mid z=0)$ requires a mean vector and a covariance matrix. For m variables, the mean vector has m parameters. The covariance is a $m \times m$ matrix. However, the matrix is symmetric so, we only need to count the diagonal and upper diagonal part of the matrix, i.e. $m + m\frac{(m-1)}{2}$. In this context, the total number of parameters is $2\left(m + m\frac{(m+1)}{2}\right) + 1$.

Considering the naïve Bayes: we need to estimate $p(y_i \mid z = c)$, requiring the fitting of a (univariate) Gaussian distribution with two parameters: μ_i and σ_i. Since there are 2 classes and m features, we have $2 \times m \times 2 = 4m$ parameters for the likelihoods. The total number of parameters is $1 + 4m$.

Programming quests

Resources: *Classification* and *Evaluation* notebooks available at the course's webpage

1. Reuse the **sklearn** code from last lab where we learnt a decision tree in the *breast.w* data:

   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** (https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_rel.html), assess whether the differences in predictive accuracy are statistically significant.