University of Southern Denmark
IMADA

DM566: Data Mining and Machine Learning
Spring term 2022

Exercise 7: Bayes Optimal Classifier, Naïve Bayes, Random Variables and Distributions, EM Clustering

Exercise 7-1 Bayes Optimal (1 point)
We have a classification problem with two classes " + " and " − ", three trained classifiers h_1, h_2, and h_3, with the following probabilities of the classifiers, given the training data D:

$$\Pr(h_1|D) = 0.4$$
$$\Pr(h_2|D) = 0.6$$
$$\Pr(h_3|D) = 0.3$$

For the three test instances o_1, o_2, o_3, the classifiers give the following class probabilities:

- o_1:

$$\Pr(+|h_1) = 0.6 \qquad \Pr(-|h_1) = 0.4$$
$$\Pr(+|h_2) = 1.0 \qquad \Pr(-|h_2) = 0.0$$
$$\Pr(+|h_3) = 0.2 \qquad \Pr(-|h_3) = 0.8$$

- o_2:

$$\Pr(+|h_1) = 0.8 \qquad \Pr(-|h_1) = 0.2$$
$$\Pr(+|h_2) = 0.6 \qquad \Pr(-|h_2) = 0.4$$
$$\Pr(+|h_3) = 0.1 \qquad \Pr(-|h_3) = 0.9$$

- o_3:

$$\Pr(+|h_1) = 0.3 \qquad \Pr(-|h_1) = 0.7$$
$$\Pr(+|h_2) = 0.1 \qquad \Pr(-|h_2) = 0.9$$
$$\Pr(+|h_3) = 1.0 \qquad \Pr(-|h_3) = 0.0$$

We combine the three classifiers to get a Bayes optimal classifier. Which class probabilities will we get from this Bayes optimal classifier for the three test instances?

Suggested solution:
The Bayes optimal classifier adds the conditional class probabilities given the classifier, weighted with the conditional classifier probabilities given the data:

$$\Pr(c_j|D) = \sum_{h_i \in H} \Pr(c_j|h_i)\Pr(h_i|D)$$

The resulting probabilities are:

$$\begin{aligned}
o_1: &\quad \Pr(+|\text{Bayes optimal}) = (0.6 \cdot 0.4 + 1 \cdot 0.6 + 0.2 \cdot 0.3) = 0.9 \\
o_1: &\quad \Pr(-|\text{Bayes optimal}) = (0.4 \cdot 0.4 + 0.0 \cdot 0.6 + 0.8 \cdot 0.3) = 0.4 \\
o_2: &\quad \Pr(+|\text{Bayes optimal}) = (0.8 \cdot 0.4 + 0.6 \cdot 0.6 + 0.1 \cdot 0.3) = 0.71 \\
o_2: &\quad \Pr(-|\text{Bayes optimal}) = (0.2 \cdot 0.4 + 0.4 \cdot 0.6 + 0.9 \cdot 0.3) = 0.59 \\
o_3: &\quad \Pr(+|\text{Bayes optimal}) = (0.3 \cdot 0.4 + 0.1 \cdot 0.6 + 1 \cdot 0.3) = 0.48 \\
o_3: &\quad \Pr(-|\text{Bayes optimal}) = (0.7 \cdot 0.4 + 0.9 \cdot 0.6 + 0 \cdot 0.3) = 0.82
\end{aligned}$$

The predictions are therefore:

$$\begin{aligned}
o_1 &: + \\
o_2 &: + \\
o_3 &: -
\end{aligned}$$

Exercise 7-2 Naïve Bayes (1 point)

The skiing season is open. To reliably decide when to go skiing and when not, you could use a classifier such as Naïve Bayes. The classifier will be trained with your observations from the last year. Your notes include the following attributes:

The weather: The attribute weather can have the following three values: sunny, rainy, and snow.

The snow level: The attribute snow level can have the following two values: ≥ 50 (There are at least 50 cm of snow) and < 50 (There are less than 50 cm of snow).

Assume you went skiing 8 times during the previous year. Here is the table with your decisions:

Weather	Snow level	Ski?
sunny	< 50	no
rainy	< 50	no
rainy	≥ 50	no
snow	≥ 50	yes
snow	< 50	no
sunny	≥ 50	yes
snow	≥ 50	yes
rainy	< 50	yes

1. Compute the *a priori* probabilities for both classes ski = yes and ski = no (on the training set)!

   Suggested solution:

$$P(ski) = 0.5$$
$$P(\neg ski) = 0.5$$

2. Compute the distribution of the conditional probabilities for the two classes for each attribute.

   Suggested solution:

$$P(weather = sunny|ski) = \frac{1}{4}$$
$$P(weather = snow|ski) = \frac{2}{4}$$
$$P(weather = rainy|ski) = \frac{1}{4}$$
$$P(weather = sunny|\neg ski) = \frac{1}{4}$$
$$P(weather = snow|\neg ski) = \frac{1}{4}$$
$$P(weather = rainy|\neg ski) = \frac{2}{4}$$

$$P(snow \geq 50|ski) = \frac{3}{4}$$
$$P(snow < 50|ski) = \frac{1}{4}$$
$$P(snow \geq 50|\neg ski) = \frac{1}{4}$$
$$P(snow < 50|\neg ski) = \frac{3}{4}$$

3. Decide for the following weather and snow conditions, whether to go skiing or not! Use the Naïve Bayes classifier as trained in the previous steps for your decision.

	Weather	Snow level
day A	sunny	≥ 50
day B	rainy	< 50
day C	snow	< 50

Suggested solution:

$$\begin{aligned} P(C_i|M) &= \frac{P(M|C_i) \cdot P(C_i)}{P(M)} \\ &= \frac{P(M|C_i) \cdot P(C_i)}{\sum_{C_j \in C} P(C_j) \cdot P(M|C_j)} \end{aligned}$$

Day A:

$$\begin{aligned} P(ski|weather = sunny, snow \geq 50) &= \frac{P(weather = sunny|ski) \cdot P(snow \geq 50|ski) \cdot P(ski)}{P(weather = sunny, snow \geq 50)} \\ &= \frac{\frac{1}{4} \cdot \frac{3}{4} \cdot \frac{1}{2}}{P(weather = sunny, snow \geq 50)} \\ &= \frac{\frac{3}{32}}{P(weather = sunny, snow \geq 50)} \end{aligned}$$

$$\begin{aligned} P(\neg ski|weather = sunny, snow \geq 50) &= \frac{P(weather = sunny|\neg ski) \cdot P(snow \geq 50|\neg ski) \cdot P(\neg ski)}{P(weather = sunny, snow \geq 50)} \\ &= \frac{\frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{2}}{P(weather = sunny, snow \geq 50)} \\ &= \frac{\frac{1}{32}}{P(weather = sunny, snow \geq 50)} \end{aligned}$$

$\Rightarrow$ Ski

Day B:

$$P(ski|weather = rainy, snow < 50) = \frac{P(weather = rainy|ski) \cdot P(snow < 50|ski) \cdot P(ski)}{P(weather = rainy, snow < 50)}$$
$$= \frac{\frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{2}}{P(weather = rainy, snow < 50)}$$
$$= \frac{\frac{1}{32}}{P(weather = rainy, snow < 50)}$$

$$P(\neg ski|weather = rainy, snow < 50) = \frac{P(weather = rainy|\neg ski) \cdot P(snow < 50|\neg ski) \cdot P(\neg ski)}{P(weather = rainy, snow < 50)}$$
$$= \frac{\frac{2}{4} \cdot \frac{3}{4} \cdot \frac{1}{2}}{P(weather = rainy, snow < 50)}$$
$$= \frac{\frac{6}{32}}{P(weather = rainy, snow < 50)}$$

$\Rightarrow$ Do not ski

Day C:

$$P(ski|weather = snow, snow < 50) = \frac{P(weather = snow|ski) \cdot P(snow < 50|ski) \cdot P(ski)}{P(weather = snow, snow < 50)}$$
$$= \frac{\frac{2}{4} \cdot \frac{1}{4} \cdot \frac{1}{2}}{P(weather = snow, snow < 50)}$$
$$= \frac{\frac{2}{32}}{P(weather = snow, snow < 50)}$$

$$P(\neg ski|weather = snow, snow < 50) = \frac{P(weather = snow|\neg ski) \cdot P(snow < 50|\neg ski) \cdot P(\neg ski)}{P(weather = snow, snow < 50)}$$
$$= \frac{\frac{1}{4} \cdot \frac{3}{4} \cdot \frac{1}{2}}{P(weather = snow, snow < 50)}$$
$$= \frac{\frac{3}{32}}{P(weather = snow, snow < 50)}$$

$\Rightarrow$ Do not ski

Exercise 7-3 Random Variables and Probability Distributions (1 point)

We played a lot with dice in the lecture. When we take the sum of n dice (a random variable), we get a probability distribution over the possible values. For just one die, this distribution is discrete with equal probabilities over $\{1, 2, 3, 4, 5, 6\}$. For two dies, the probabilities are unequally distributed over $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$.

How does the shape of the probability distribution develop with increasing n?

Visualize the development of the shape with a little program or script.

Suggested solution:

Can be done with a simple python script, using `matplotlib`.

```
import matplotlib.pyplot as plt

# one die
probabilities = [(1/6)]*6
sums = [1, 2, 3, 4, 5, 6]
plt.scatter(sums, probabilities)
plt.show()

# two dice
maxsum = 6*2
probabilities = [0]*maxsum
sums = range(1,maxsum+1)
for i in range(1, 6):
  for j in range(1, 6):
    s = i+j
    probabilities[s] += (1/maxsum)
plt.scatter(sums, probabilities)
plt.show()

# three dice
maxsum = 6*3
probabilities = [0]*maxsum
sums = range(1,maxsum+1)
for i in range(1, 6):
  for j in range(1, 6):
    for k in range(1, 6):
      s = i+j+k
      probabilities[s] += (1/maxsum)
plt.scatter(sums, probabilities)
plt.show()

# four dice
maxsum = 6*4
probabilities = [0]*maxsum
```

```
sums = range(1,maxsum+1)
for i in range(1, 6):
  for j in range(1, 6):
    for k in range(1, 6):
      for l in range(1, 6):
        s = i+j+k+l
        probabilities[s] += (1/maxsum)
plt.scatter(sums, probabilities)
plt.show()

# five dice
maxsum = 6*5
probabilities = [0]*maxsum
sums = range(1,maxsum+1)
for i in range(1, 6):
  for j in range(1, 6):
    for k in range(1, 6):
      for l in range(1, 6):
        for m in range(1, 6):
          s = i+j+k+l+m
          probabilities[s] += (1/maxsum)
plt.scatter(sums, probabilities)
plt.show()

# six dice
maxsum = 6*6
probabilities = [0]*maxsum
sums = range(1,maxsum+1)
for i in range(1, 6):
  for j in range(1, 6):
    for k in range(1, 6):
      for l in range(1, 6):
        for m in range(1, 6):
          for o in range(1, 6):
            s = i+j+k+l+m+o
            probabilities[s] += (1/maxsum)
plt.scatter(sums, probabilities)
plt.show()
```

Exercise 7-4 Assignments in the EM-Algorithm (1 point)

Given a data set with 100 points consisting of three Gaussian clusters A, B and C and the point p.

The cluster A contains 30% of all objects and is represented using the mean of all its points $\mu_A = (2,2)$ and the covariance matrix $\Sigma_A = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}$.

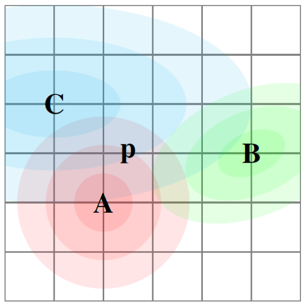
You will need the inverse $\Sigma_A^{-1} = \begin{pmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{3} \end{pmatrix}$.

The cluster B contains 20% of all objects and is represented using the mean of all its points $\mu_B = (5,3)$ and the covariance matrix $\Sigma_B = \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}$. $\Sigma_B^{-1} \approx \begin{pmatrix} 0.571428 & -0.142857 \\ -0.142857 & 0.285714 \end{pmatrix}$.

The cluster C contains 50% of all objects and is represented using the mean of all its points $\mu_C = (1,4)$ and the covariance matrix $\Sigma_C = \begin{pmatrix} 16 & 0 \\ 0 & 4 \end{pmatrix}$. $\Sigma_C^{-1} = \begin{pmatrix} \frac{1}{16} & 0 \\ 0 & \frac{1}{4} \end{pmatrix}$.

The point p is given by the coordinates $(2.5, 3.0)$.

The following sketch is not exact, and only gives a rough idea of the cluster locations:



Compute the three probabilities of p belonging to the clusters A, B, and C.

Suggested solution: Cluster A:

$$\Sigma_A = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix} \quad \Sigma_A^{-1} = \begin{pmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{3} \end{pmatrix} \quad \mu_A = (2,2)$$

$$p - \mu_A = (0.5, 1)$$

$$dist^2 = (p-\mu)^T \Sigma^{-1} (p-\mu) \approx 0.41666$$

$$dens_A \approx \frac{1}{\sqrt{(2\pi)^2 9}} e^{-\frac{1}{2} 0.41666}$$
$$\approx 0.04307456$$

Cluster B:

$$\Sigma_B = \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix} \quad \Sigma_B^{-1} \approx \begin{pmatrix} 0.571428 & -0.142857 \\ -0.142857 & 0.285714 \end{pmatrix} \quad \mu_B = (5,3)$$

$$p - \mu_B = (-2.5, 0)$$

$$dist^2 = (p-\mu)^T \Sigma^{-1} (p-\mu) \approx 3.5714285$$

$$dens_B \approx \frac{1}{\sqrt{(2\pi)^2 7}} e^{-\frac{1}{2} 3.5714285}$$
$$\approx 0.01008661$$

Cluster C:

$$\Sigma_C = \begin{pmatrix} 16 & 0 \\ 0 & 4 \end{pmatrix} \quad \Sigma_C^{-1} = \begin{pmatrix} \frac{1}{16} & 0 \\ 0 & \frac{1}{4} \end{pmatrix} \quad \mu_C = (1,4)$$

$$p - \mu_C = (1.5, -1)$$

$$dist^2 = (p-\mu)^T \Sigma^{-1} (p-\mu) \approx 0.390625$$

$$dens_C \approx \frac{1}{\sqrt{(2\pi)^2 64}} e^{-\frac{1}{2} 0.390625}$$
$$\approx 0.01636466$$

	A	B	C
Density	0.043075	0.010087	0.016365
Size	30%	20%	50%
Score	0.012922	0.002017	0.008182
Sum	$\frac{0.012922}{0.023122}$	$\frac{0.002017}{0.023122}$	$\frac{0.008182}{0.023122}$
Weight	55.9%	8.2%	35.4%