Практические задания №7. Григорьев И.С. 6304

Задание №1

Даны следующие данные

$\mathbf{x}_i$	a_1	a_2	a_3	Class
$\mathbf{x}_1$	T	T	5.0	Y
$\mathbf{x}_2$	T	T	7.0	Y
$\mathbf{x}_3$	T	F	8.0	N
$\mathbf{x}_4$	F	F	3.0	Y
$\mathbf{x}_5$	F	T	7.0	N
$\mathbf{x}_6$	F	T	4.0	N
$\mathbf{x}_7$	F	F	5.0	N
$\mathbf{x}_8$	T	F	6.0	Y
$\mathbf{x}_9$	F	T	1.0	N

Используя наивный байесовский классификатор определите класс точки $\boldsymbol{x}$=(T,F,1.0).

Исходя из предположения, что все атрибуты независимы $P(\boldsymbol{x}|c_i)$ может быть разложена на произведение вероятностей каждого измерения:

$$P(\boldsymbol{x}|c_i) = \prod_{j=1}^{d} P(x_j|c_i)$$

Для числовых атрибутов:

$$P(x_j|c_i) \propto \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp\left\{-\frac{(x_j-\mu_{ij})^2}{2\sigma_{ij}^2}\right\}$$

Для категориальных атрибутов:

$$\prod_{j=1}^{d} P(x_j|c_i) = \prod_{j=1}^{d} \frac{n_i(\mathbf{v}_j)}{n_i}$$

где $n_i(\mathbf{v}_j)$ частота категориального значения определенного атрибута и класса.

```
a3_Y = np.array([5, 7, 3, 6])
print('a3_Y mean = ', a3_Y.mean())
print('a3_Y std = ', a3_Y.std())
```

```
a3_Y mean =  5.25
a3_Y std =  1.479019945774904
```

```
a3_N = np.array([8, 7, 4, 5, 1])
print('a3_N mean = ', a3_N.mean())
print('a3_N std = ', a3_N.std())
```

```
a3_N mean =  5.0
a3_N std =  2.449489742783178
```

```
def f(x, mean, std):
    return np.exp(- ((x - mean) ** 2) / (2 * std**2)) / (np.sqrt(2 * np.pi) *␣
 ↪std)
```

```
P1_Y = f(1.0, a3_Y.mean(), a3_Y.std())
P1_Y
```

```
0.0043443043728038045
```

```
P1_N = f(1.0, a3_N.mean(), a3_N.std())
P1_N
```

```
0.04293140793792167
```

```
P_Y_v = P1_Y * (3/4) * (1/2) * (4/9)
P_Y_v
```

```
0.0007240507288006341
```

```
P_N_v = P1_N * (1/5) * (2/5) * (5/9)
P_N_v
```

```
0.0019080625750187413
```

$P(N|\boldsymbol{x}) > P(Y|\boldsymbol{x})$, значит **x** = (T,F,1.0) относится к классу N

Задание №2

Даны два класса c_1 и c_2 со следующими мат. ожиданиями и матрицами ковариации:

$$\boldsymbol{\mu}_1 = (1,3) \qquad \boldsymbol{\mu}_2 = (5,5)$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 5 & 3 \\ 3 & 2 \end{pmatrix} \qquad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$$

Классифицируйте точку **x** = (3,4) используя Байесовский вывод, предположив, что классы распределены по нормальному закону, и P(*c1*) = P(*c2*) = 0.5

Т.к. классы нормально распределены, плотность вероятности в x для класса c_i задается как

$$f_i(\boldsymbol{x}) = f(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(\sqrt{2\pi})^d\sqrt{|\boldsymbol{\Sigma}_i|}} \exp\left\{-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)}{2}\right\}$$

Вероятность наблюдения **x** при условии, что истинным классом является c_i, при рассмотрении малого интервала $\epsilon > 0$ можно получить как

$$P(\boldsymbol{x}|c_i) = 2\epsilon \cdot f_i(\boldsymbol{x})$$

Тогда апостериорную вероятность можно получить по теореме Байеса как

$$P(c_i|\boldsymbol{x}) = \frac{2\epsilon \cdot f_i(\boldsymbol{x})P(c_i)}{\sum_{j=1}^{k} 2\epsilon \cdot f_j(\boldsymbol{x})P(c_j)} = \frac{f_i(\boldsymbol{x})P(c_i)}{\sum_{j=1}^{k} f_j(\boldsymbol{x})P(c_j)}$$

Таким образом предсказать класс для $\boldsymbol{x}$ можно следующим образом

$$\hat{y} = \arg\max_{c_i}\{f_i(\boldsymbol{x})P(c_i)\}$$

```
def f(x, m, cov):
    return np.exp(- (np.dot(np.dot((x - m), np.linalg.inv(cov)), (x - m))) / 2)␣
 ↪/ (2 * np.pi * np.sqrt(np.linalg.det(cov)))
```

```
p1 = f(np.array([3, 4]), np.array([1, 3]), np.array([[5, 3], [3, 2]])) * 0.5
p1
```

```
0.04826617631502697
```

```
p2 = f(np.array([3, 4]), np.array([5, 5]), np.array([[2, 0], [0, 1]])) * 0.5
p2
```

```
0.012555482738023717
```

$P(c_1|\boldsymbol{x}) > P(c_2|\boldsymbol{x})$, значит **x** = (3,4) относится к первому классу c_1

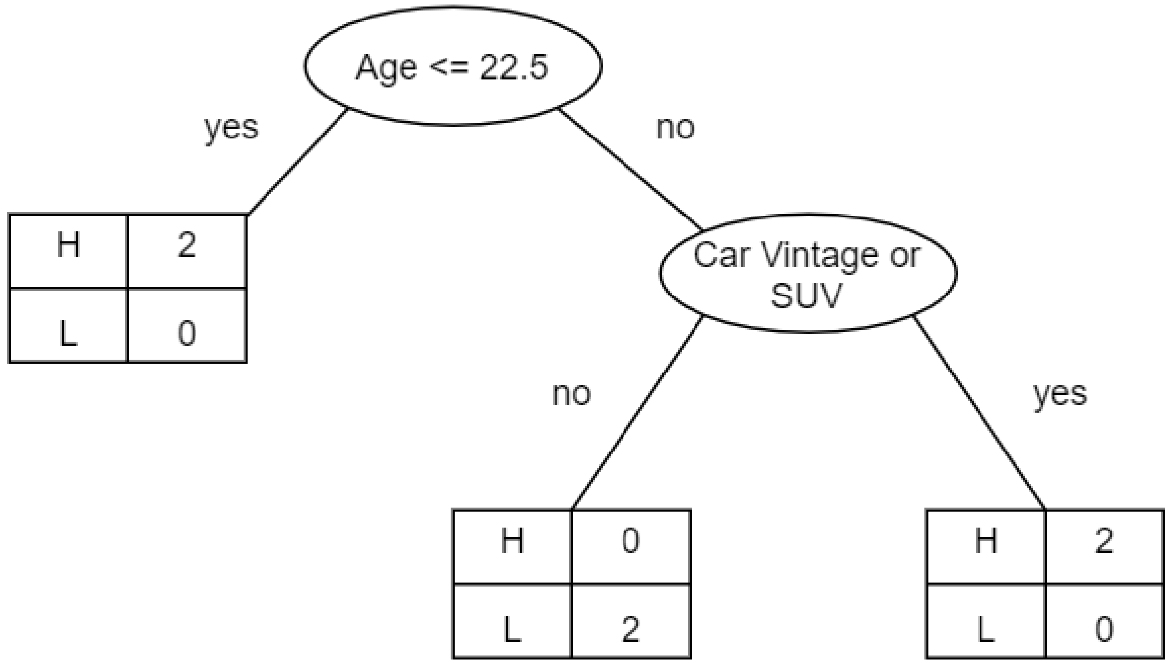
Задание №3

Даны следующие данные

Point	Age	Car	**Risk**
$\mathbf{x}_1$	25	Sports	*L*
$\mathbf{x}_2$	20	Vintage	*H*
$\mathbf{x}_3$	25	Sports	*L*
$\mathbf{x}_4$	45	SUV	*H*
$\mathbf{x}_5$	20	Sports	*H*
$\mathbf{x}_6$	25	SUV	*H*

Постройте решающее дерево используя порог для чистоты (purity threshold) равным 100%.

В качестве критерия для разделения используйте энтропию. Классифицируйте наблюдение (Age=27, Car=Vintage)



Age 27 <= 22.5 => no => Car Vintage or SUV? => yes => class H