Statistical Machine Learning

Lecture 4: Classification

2022-23

Mahalanobis Distance

- Given a vector space $\mathcal{V}$ a distance d is an application from $\mathcal{V} \times \mathcal{V}$ into the non-negative rals $\mathbb{R}^+ \cup \{0\}$ such that:
  1. For all $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathcal{V}$ $d(\boldsymbol{x}, \boldsymbol{y}) \geq 0$ and equal to 0 if and only if $\boldsymbol{x} = \boldsymbol{y}$. (Positivity)
  2. For all $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathcal{V}$ $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$. (Simetry)
  3. For all $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$ in $\mathcal{V}$ $d(\boldsymbol{x}, \boldsymbol{y}) \leq d(\boldsymbol{x}, \boldsymbol{z}) + d(\boldsymbol{z}, \boldsymbol{y})$. (Triangula Inequality)

Mahalanobis Distance

- Gaussian distribution:

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

$$\Delta^2 \equiv (\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$$

  where Δ is the Mahalanobis distance determined by the matrix $\boldsymbol{\Sigma}$.
- $\boldsymbol{\Sigma}$ is positive definite, therefore the solutions to the eigenvalue problem:

$$\boldsymbol{\Sigma}\boldsymbol{u}_\lambda = \lambda\boldsymbol{u}_\lambda$$

  are such that $\lambda \in \mathbb{R}^+$ and for any pair $\lambda \neq \lambda'$, $\boldsymbol{u}_\lambda^T\boldsymbol{u}_{\lambda'} = 0$.

Mahalanobis Distance

- Let us define the matrix $\boldsymbol{U} = \left(\boldsymbol{u}_{\lambda_1}, \boldsymbol{u}_{\lambda_2}, \ldots, \boldsymbol{u}_{\lambda_d}\right)$. By the properties of the eigenvectors, $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^T = \mathbf{1}$.
- Let us define the matrix $\boldsymbol{\Lambda}$ such that $[\boldsymbol{\Lambda}]_{i,i} = \lambda_i$ and $[\boldsymbol{\Lambda}]_{i,j} = 0$ for all $i \neq j$.
- Observe that $[\boldsymbol{\Lambda}, \boldsymbol{U}] = \boldsymbol{\Lambda}\boldsymbol{U} - \boldsymbol{U}\boldsymbol{\Lambda} = \mathbf{0}$.
- Then

$$
\begin{aligned}
\boldsymbol{\Sigma}\boldsymbol{U} &= \boldsymbol{\Lambda}\boldsymbol{U} = \boldsymbol{U}\boldsymbol{\Lambda} \\
\boldsymbol{U}^T\boldsymbol{\Sigma}\boldsymbol{U} &= \boldsymbol{\Lambda} \\
\boldsymbol{\Sigma} &= \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T \\
\boldsymbol{\Sigma}^{-1} &= \left(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T\right)^{-1} = \left(\boldsymbol{U}^T\right)^{-1}(\boldsymbol{U}\boldsymbol{\Lambda})^{-1} \\
&= \boldsymbol{U}\boldsymbol{\Lambda}^{-1}(\boldsymbol{U})^{-1} = \boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^T
\end{aligned}
$$

Mahalanobis Distance

- For any given pair of vectors $\boldsymbol{x}$, $\boldsymbol{y} \in \mathbb{R}^d$ and a covariance matrix $\boldsymbol{\Sigma}$ we have that:

$$\begin{aligned}
\Delta^2(\boldsymbol{x}, \boldsymbol{y}) &= (\boldsymbol{x} - \boldsymbol{y})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{y}) \\
&= (\boldsymbol{x} - \boldsymbol{y})^T \boldsymbol{U} \boldsymbol{\Lambda}^{-1} \boldsymbol{U}^T (\boldsymbol{x} - \boldsymbol{y}) \\
&= \left[\boldsymbol{U}^T (\boldsymbol{x} - \boldsymbol{y})\right]^T \boldsymbol{\Lambda}^{-1} \boldsymbol{U}^T (\boldsymbol{x} - \boldsymbol{y}) \\
&= \sum_{i=1}^{d} \frac{\left(\left[\boldsymbol{U}^T (\boldsymbol{x} - \boldsymbol{y})\right]_i\right)^2}{\lambda_i},
\end{aligned}$$

similar to the Euclidian distance.

Mahalanobis Distance

- Given the vectors $\boldsymbol{r}^T = (x, y)$ and $\boldsymbol{\mu} = (1, 1)$ find the set that satisfies $\Delta^2(\boldsymbol{r}, \boldsymbol{\mu}) = 4$, for a covariance $\boldsymbol{\Sigma} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$.
- Eigenvalues : First observe that the eigenvalues of the covariance satisfy the quadratic equation: $(2-\lambda)^2 - 1 = 0$ therefore the solutions are $\lambda = 1$ and $\lambda = 3$.
- Eigenvectors:

$$\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \lambda \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \quad \text{or} \quad u_2 = (2-\lambda)u_1,$$

  thus, for $\lambda = 1$ we have that $u_1 = u_2$ and for $\lambda = 3$ we have that $u_1 = -u_2$.
- The matrix of rotation $\boldsymbol{U}$ becomes

$$\boldsymbol{U} = \frac{\sqrt{2}}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \boldsymbol{U}^T.$$

Mahalanobis Distance

- The covariance matrix can be expressed as:

$$\mathbf{\Sigma} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

$$\mathbf{\Lambda} = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$$

$$\mathbf{\Sigma}^{-1} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{3} \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

- The distance becomes:

$$\begin{aligned} 4 &= \Delta^2(\boldsymbol{r}, \boldsymbol{\mu}) \\ &= (x-1, y-1)\frac{\sqrt{2}}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \times \\ &\times \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{3} \end{pmatrix}\frac{\sqrt{2}}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} x-1 \\ y-1 \end{pmatrix} \end{aligned}$$

Mahalanobis Distance

- The distance becomes:

$$
\begin{aligned}
4 &= \Delta^2(\boldsymbol{r}, \boldsymbol{\mu}) \\
&= \left[\frac{\sqrt{2}}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x-1 \\ y-1 \end{pmatrix} \right]^T \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{3} \end{pmatrix} \times \\
&\quad \times \left[\frac{\sqrt{2}}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x-1 \\ y-1 \end{pmatrix} \right].
\end{aligned}
$$

- Let us define the new shifted (by $\boldsymbol{\mu}$) and rotated (by $\boldsymbol{U}$) vector $\boldsymbol{r}'$:

$$
\boldsymbol{r}' = \begin{pmatrix} x' \\ y' \end{pmatrix} = \frac{\sqrt{2}}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x-1 \\ y-1 \end{pmatrix} = \begin{pmatrix} \frac{(x-1)+(y-1)}{\sqrt{2}} \\ \frac{(x-1)-(y-1)}{\sqrt{2}} \end{pmatrix},
$$

thus

$$x' = \frac{(x-1)+(y-1)}{\sqrt{2}}$$
$$y' = \frac{(x-1)-(y-1)}{\sqrt{2}},$$

Mahalanobis Distance

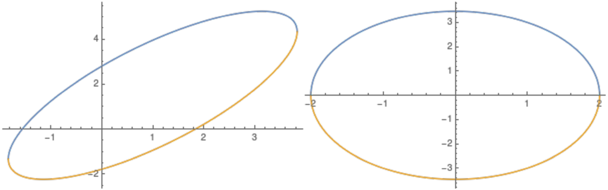
- The distance becomes:

$$4 = \Delta^2(\boldsymbol{r}, \boldsymbol{\mu}) = (x', y') \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{3} \end{pmatrix} \begin{pmatrix} x' \\ y' \end{pmatrix}$$

$$4 = (x')^2 + \frac{(y')^2}{3}$$

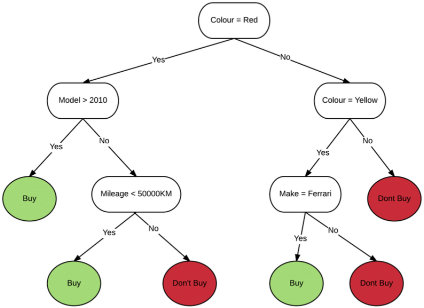
$$1 = \frac{(x')^2}{4} + \frac{(y')^2}{12},$$

which is the equation of an ellipsis centered at $\mathbf{0}$ with axis $a = 2$ and $b = \sqrt{12}$.



Decision Trees

1. Supervised learning.
2. Determines a course of action.
3. Each branch represents a possible decision.
4. Descriptive on how are decisions are taken.



Definitions

1. There are three type of nodes in a decision tree: root, decision and leaf.
2. Root node is where the entire population of the data set sits before any decision is taken.
3. Decision node is where a split in the sample is performed according to an action.
4. Leaf node is a terminal node where no more decisions are taken and a final classification is reached.

Entropy

1. Given $p(x)$ the information content of outcome x is defined as $h(x) = -\log_2 p(x)$.
2. The Entropy of a probability distribution is defined as the expected information content $H[p] = \sum_{x \in \mathcal{V}} p(x)h(x)$.
3. $H[p] \geq 0$ with $=$ if and only if there exists $x_0 \in \mathcal{V}$ such that $p(x_0) = 1$.
4. $H[p]$ is maximized if $p(x) = p(x')$ for any pair $x, x' \in \mathcal{V}$.

Information-Based Decisions

- Let as assume we have features $\boldsymbol{x} \in \mathcal{V}^d$ and labels $t \in \mathcal{T}$ (most probably $\{\pm 1\}$).
- Let as assume that there exists a probability $p(t, \boldsymbol{x})$ such that

$$p_t(t) = \prod_{j=1}^{d} \sum_{x_j \in \mathcal{V}} p(t, \boldsymbol{x})$$

$$p_{t,\ell}(t, x) = \prod_{j \neq \ell} \sum_{x_j \in \mathcal{V}} p(t, \boldsymbol{x})$$

$$p_\ell(x) = \sum_{t \in \mathcal{T}} p_{t,\ell}(t, x)$$

$$p_{t|\ell}(t|x) = \frac{p_{t,\ell}(t, x)}{p_\ell(x)}$$

Information-Based Decisions

- Given the marginal (prior) probability $p(t)$ the Entropy associated with it is defined as the functional:

$$H[p] \equiv -\sum_{t \in \mathcal{T}} p_t(t) \log_2 p_t(t).$$

- On the same path we can define the conditional entropy function:

$$h_\ell(x) \equiv -\sum_{t \in \mathcal{T}} p_{t|\ell}(t|x) \log_2 p_{t|\ell}(t|x)$$

and the associated functional:

$$H_\ell[p] = \sum_{x \in \mathcal{V}} p_\ell(x) h_\ell(x)$$

Information-Based Decisions

- Information Gain associated to the ℓ-th feature x_ℓ:

$$I_\ell[p] \equiv H[p] - H_\ell[p],$$

- Splits are done according to the feature x_ℓ that provides the largest Information Gain.

Example

- Suppose we have the following data set

	x_1	x_2	x_3	t
0	1	1	1	1
1	1	1	1	1
2	1	1	0	0
3	0	1	1	1
4	1	1	1	1
5	1	1	1	1
6	1	0	0	0
7	1	1	0	0
8	1	1	1	1
9	0	1	1	0

Example

- Let us consider first $p(t)$ and $p(x_\ell)$

$$
\begin{aligned}
p(t=1) &= \frac{6}{10} & \qquad p(t=0) &= \frac{4}{10}, \\
p(x_1=1) &= \frac{8}{10} & \qquad p(x_1=0) &= \frac{2}{10}, \\
p(x_2=1) &= \frac{9}{10} & \qquad p(x_2=0) &= \frac{1}{10}, \\
p(x_3=1) &= \frac{7}{10} & \qquad p(x_3=0) &= \frac{3}{10},
\end{aligned}
$$

Example

- and the other marginals

$$p(t=1|x_1=1)=\frac{5}{8} \qquad p(t=0|x_1=1)=\frac{3}{8}$$
$$p(t=1x_1=0)=\frac{1}{2} \qquad p(t=0|x_1=0)=\frac{1}{2}$$

$$p(t=1|x_2=1)=\frac{6}{9} \qquad p(t=0|x_2=1)=\frac{3}{9}$$
$$p(t=1|x_2=0)=0 \qquad p(t=0|x_2=0)=1$$

$$p(t=1|x_3=1)=\frac{6}{7} \qquad p(t=0|x_3=1)=\frac{1}{7}$$
$$p(t=1|x_3=0)=0 \qquad p(t=0|x_3=0)=1$$

Example

- Let us compute the entropy

$$H[p] = -\left[\frac{6}{10}\log_2\frac{6}{10} + \frac{4}{10}\log_2\frac{4}{10}\right] = 0.971$$

- and $H_\ell[p]$

$$\begin{aligned}
H_1[p] &= -\left\{\frac{8}{10}\left[\frac{5}{8}\log_2\frac{5}{8} + \frac{3}{8}\log_2\frac{3}{8}\right] + \frac{2}{10}\left[\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right]\right\} \\
&= 0.964 \\
H_2[p] &= -\left\{\frac{9}{10}\left[\frac{6}{9}\log_2\frac{6}{9} + \frac{3}{9}\log_2\frac{3}{9}\right] + \frac{1}{10}\left[0\log_2 0 + 1\log_2 1\right]\right\} \\
&= 0.826 \\
H_3[p] &= -\left\{\frac{7}{10}\left[\frac{6}{7}\log_2\frac{6}{7} + \frac{1}{7}\log_2\frac{1}{7}\right] + \frac{3}{10}\left[0\log_2 0 + 1\log_2 1\right]\right\} \\
&= 0.414
\end{aligned}$$

Example

- and then

$$
\begin{aligned}
I_1[p] &= 0.007 \\
I_2[p] &= 0.145 \\
I_3[p] &= 0.557
\end{aligned}
$$

Example

- Since the largest information gain is associated to x_3 we use this feature to split the data:

	x_1	x_2	$x_3 = 1$	t
0	1	1	1	1
1	1	1	1	1
3	0	1	1	1
4	1	1	1	1
5	1	1	1	1
8	1	1	1	1
9	0	1	1	0

	x_1	x_2	$x_3 = 0$	t
2	1	1	0	0
6	1	0	0	0
7	1	1	0	0

Example

- Let us consider first $p(t, x_3 = 1)$ and $p(x_{\ell \neq 3}, x_3 = 1)$

$$
\begin{aligned}
p(t = 1 | x_3 = 1) &= \frac{6}{7} & \qquad p(t = 0 | x_3 = 1) &= \frac{1}{7}, \\
p(x_1 = 1 | x_3 = 1) &= \frac{5}{7} & \qquad p(x_1 = 0 | x_3 = 1) &= \frac{2}{7}, \\
p(x_2 = 1 | x_3 = 1) &= 1 & \qquad p(x_2 = 0 | x_3 = 1) &= 0,
\end{aligned}
$$

Example

- and the other marginals

$$p(t=1|x_1=1,x_3=1)=1 \qquad p(t=0|x_1=1,x_3=1)=0$$

$$p(t=1|x_1=0,x_3=1)=\frac{1}{2} \qquad p(t=0|x_1=0,x_3=1)=\frac{1}{2}$$

$$p(t=1|x_2=1,x_3=1)=\frac{6}{7} \qquad p(t=0|x_2=1,x_3=1)=\frac{1}{7}$$

$$p(t=1|x_2=0,x_3=1)=0 \qquad p(t=0|x_2=0,x_3=1)=0$$

observe that the event with $x_3=1$ and $x_2=0$ does not occur therefore we cannot define the probability.

Example

- Let us compute the entropy

$$H[p, x_3 = 1] = -\left[\frac{6}{7}\log_2\frac{6}{7} + \frac{1}{7}\log_2\frac{1}{7}\right] = 0.592$$

- and $H_\ell[p]$

$$\begin{aligned} H_1[p, x_3 = 1] &= -\left\{\frac{5}{7}\left[1\log_2 1 + 0\log_2 0\right] + \frac{2}{7}\left[\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right]\right\} \\ &= 0.286 \\ H_2[p, x_3 = 1] &= -\left\{1\left[\frac{6}{7}\log_2\frac{6}{7} + \frac{1}{7}\log_2\frac{1}{7}\right] + 0\,[\text{undefined}]\right\} \\ &= 0.592 \end{aligned}$$

Example

- and then

$$I_1[p, x_3 = 1] = 0.307$$
$$I_2[p, x_3 = 1] = 0$$

Example

- Since the largest information gain is associated to x_3 we use this feature to split thedata:

	$x_1 = 1$	x_2	$x_3 = 1$	t
0	1	1	1	1
1	1	1	1	1
4	1	1	1	1
5	1	1	1	1
8	1	1	1	1

	$x_1 = 0$	x_2	$x_3 = 1$	t
3	0	1	1	1
9	0	1	1	0

Example

- Final tree