CSCI567 — WA1

Jiawei Huang 7148294267

05/07/2020

1 Nearest Neighbor Classification 1

Answer: Assuming there is one point x' in the training set. A new point x is going to be predicted. if data is normalized with unit L2 norm, that is, $\|x\|_2 = \sum_{d=1}^{D} x_d^2 = 1$ for all x in the training and test sets.

- Using Euclidean distance.

$$E(x, x') = \|x - x'\|_2^2 = \sum_{d=1}^{D}(x_d - x'_d)^2 = \sum_{d=1}^{D} x_d^2 + \sum_{d=1}^{D} x_d'^2 - 2\sum_{d=1}^{D} x_d x'_d = 2(1 - \sum_{d=1}^{D} x_d x'_d)$$

- Using cosine distance.

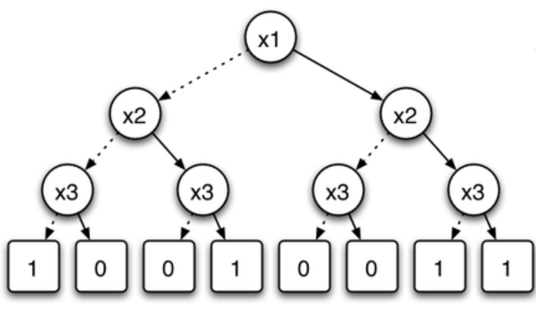
$$C(x, x') = 1 - \frac{\sum_{d=1}^{D} x_d x'_d}{\|x\|_2 \, \|x'\|_2} = 1 - \sum_{d=1}^{D} x_d x'_d = 0.5E(x, x')$$

So we have $E(x, x') = 2C(x, x')$. So changing the distance function from the Euclidean distance to the cosine distance will **NOT** affect the nearest neighbor classification results.

2 Nearest Neighbor Classification 2

1. **Answer:** We **can** have a decision tree to classify the dataset with zero classification error w.r.t. their labels if there are no conflicting data samples. conflicting data samples is two data samples $\mathbf{x}, \mathbf{x}'$ where $x_1 = x'_1, x_2 = x'_2, ..., x_{100} = x'_{100}$ but their labels are different.

   Figure 1 shows a case when dimension = 3, it is the same when the dimension is 100.



x1	x2	x3	*f*
0	0	0	1
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	0
1	1	0	1
1	1	1	1

Figura 1: a case when dimension is 3

2. **Answer:** We **can** use 1-NN to get the same result. A simple implementation is using a 100 dimension vector to cover all points from the dataset.

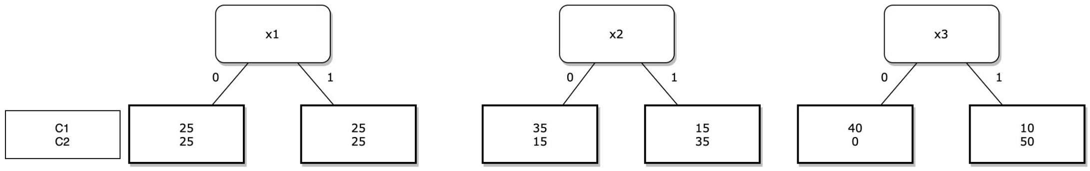


Figura 2: splitting results of different input

3 Decision Tree

1. **Answer:**

   The cases when first splitting on different inputs are shown in *Figure 2*

   - x_1: mis-classification rate:

     $$m_1 = \frac{25+25}{25+25+25+25} = \frac{1}{2}$$

     Cross entropy:
     left branch:$-(\frac{25}{25+25}log\frac{25}{25+25} + \frac{25}{25+25}log\frac{25}{25+25}) = 1$
     right branch:$-(\frac{25}{25+25}log\frac{25}{25+25} + \frac{25}{25+25}log\frac{25}{25+25}) = 1$

     $$c_1 = \frac{50}{100} + \frac{50}{100} = 1$$

     Gini index:
     left branch:$\frac{25}{25+25}(1 - \frac{25}{25+25}) + \frac{25}{25+25}(1 - \frac{25}{25+25}) = \frac{1}{2}$
     right branch:$\frac{25}{25+25}(1 - \frac{25}{25+25}) + \frac{25}{25+25}(1 - \frac{25}{25+25}) = \frac{1}{2}$

     $$g_1 = \frac{50}{100} \times \frac{1}{2} + \frac{50}{100} \times \frac{1}{2} = \frac{1}{2}$$

   - x_2:
     mis-classification rate:

     $$m_2 = \frac{15+15}{15+35+15+35} = \frac{3}{10}$$

     Cross entropy:
     left branch:$-(\frac{15}{15+35}log\frac{15}{15+35} + \frac{35}{15+35}log\frac{35}{15+35}) = 0.88$
     right branch:$-(\frac{15}{15+35}log\frac{15}{15+35} + \frac{35}{15+35}log\frac{35}{15+35}) = 0.88$

     $$c_2 = \frac{50}{100} \times 0.88 + \frac{50}{100} \times 0.88 = 0.88$$

     Gini index:
     left branch:$\frac{15}{15+35}(1 - \frac{15}{15+35}) + \frac{35}{15+35}(1 - \frac{35}{15+35}) = 0.42$
     right branch:$\frac{15}{15+35}(1 - \frac{15}{15+35}) + \frac{35}{15+35}(1 - \frac{35}{15+35}) = 0.42$

     $$g_2 = \frac{50}{100} \times 0.42 + \frac{50}{100} \times 0.42 = 0.42$$

   - x_3:
     mis-classification rate:

     $$m_3 = \frac{10}{40+0+10+50} = 0.1$$

     Cross entropy:
     left branch:$-(\frac{40}{40+0}log\frac{40}{40+0} + \frac{0}{40+0}log\frac{0}{40+0}) = 0$

right branch:$-(\frac{10}{10+50}log\frac{10}{10+50}+\frac{50}{10+50}log\frac{50}{10+50})=0.65$

$$c_3=\frac{40}{100}\times 0+\frac{60}{100}\times 0.65=0.39$$

Gini index:
left branch:$\frac{0}{40+0}(1-\frac{0}{40+0})+\frac{40}{40+0}(1-\frac{40}{40+0})=0$
right branch:$\frac{10}{10+50}(1-\frac{10}{10+50})+\frac{50}{10+50}(1-\frac{50}{10+50})=\frac{5}{18}$

$$g_3=\frac{40}{100}\times 0+\frac{60}{100}\times\frac{5}{18}=\frac{1}{6}$$

2. **Answer:**

   According to **3.1**, we know $g_3<g_2<g_1$. So we should choose x_3 to first split.

3. **Answer:**

   - start with x_1

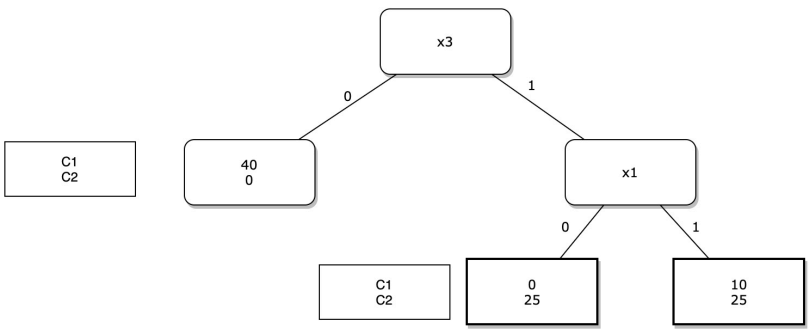


Figura 3: start with x_1

We can calculate the Gini Index in the second layer.

$$g_{x_1}=\frac{10\times 25}{35\times 35}\times 2\times\frac{35}{100}=\frac{1}{7}$$

There are **10** points incorrectly classified.(If the two branches of x1 are labeled as C1, C2, there should be 25 mis-labeled points.)

   - start with x_2

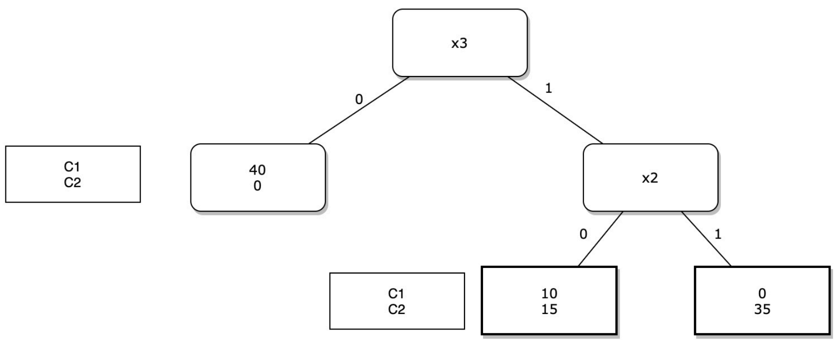


Figura 4: start with x_2

We can calculate the Gini Index in the second layer.

$$g_{x_2}=\frac{10\times 15}{25\times 25}\times 2\times\frac{25}{100}=\frac{3}{25}<g_{x_1}$$

There are **10** points incorrectly classified. (If the two branches of x2 are labeled as C1, C2, there should be 15 mis-labeled points.)

So it's better for us to choose $\mathbf{x_2}$ as the second splitting input and there are **10** points which are mis-classified. (If the two branches of x2 are labeled as C1, C2, there should be 15 mis-labeled points.)

Note: Here is showing the case of reduced error pruning. We can find out when labeling all the right branch into C2, there is only 10 mis-labeled points, which is smaller than the two-level decision tree, by doing the pruning, we get a one-level tree.

4. **Answer:**

- start with x_2

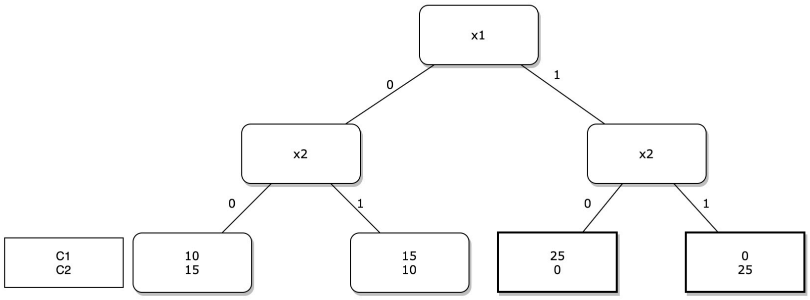


Figura 5: x_2 as second splitting layer

We can calculate the mis-classification rate in the second layer.

$$m_{x_2} = \frac{20}{100} = 0.2$$

There are **20** points incorrectly classified.

- start with x_3

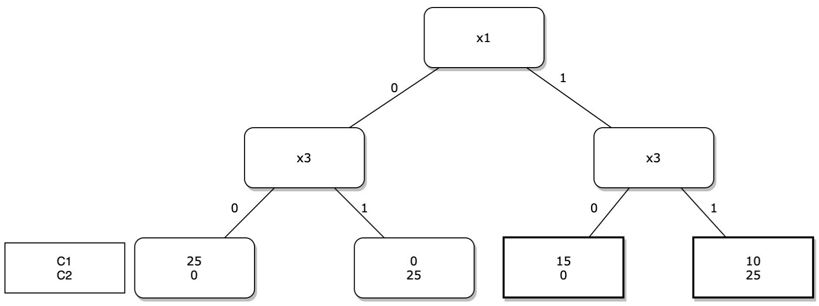


Figura 6: x_3 as second splitting layer

We can calculate the mis-classification rate in the second layer.

$$m_{x_3} = \frac{10}{100} = 0.1 < m_{x_2}$$

There are **10** points incorrectly classified.

- left with x_3 and right with x_2(see *Figure 7*)
  There are no mis-labeled points, mis-classification rate is **0**.

So we should choose $\mathbf{x_3}$ on the left and $\mathbf{x_2}$ on the right. In this way, there will be **0** mis-classified points.

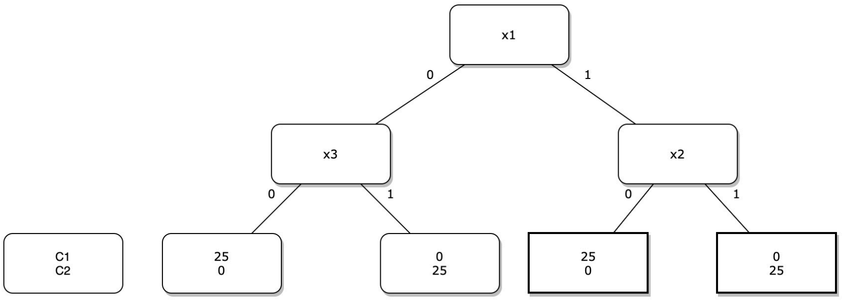


Figura 7: x_3 on the left and x_2 on the right

5. **Answer:**

   The decision tree in problem **3.4** performs better. Because after two layers, there are less points that need to be split further(it's already classified). **3.3** is a greedy method and cannot guarantee best decision tree structure.

4 Naive Bayes

1. **Answer**:

$$P(y=1|x) = \frac{P(y=1,x)}{P(x)} = \frac{P(y=1,x)}{P(y=1,x)+P(y=0,x)} = \frac{1}{1+\frac{P(y=0,x)}{P(y=1,x)}} = \frac{1}{1+\frac{P(y=0|x)}{P(y=1|x)}}$$

2. **Answer**:

$$\begin{aligned} P(y=k|x) &= \frac{P(x|y=y_k)P(y=k))}{P(x)} = \frac{1}{P(x)}exp\left[ln(\prod_{j=1}^{D} P(x_j|y=y_k)P(y=y_k))\right] \\ &= \frac{1}{P(x)}exp\left[ln(\prod_{j=1}^{D} \theta_{jk}^{x_j}(1-\theta_{jk})^{1-x_j}P(y=y_k))\right] \\ &= \frac{1}{P(x)}exp\left[\sum_{j=1}^{D} x_j ln\theta_{jk} + (1-x_j)(1-ln\theta_{jk}) + ln\pi_k\right] \\ &= \frac{1}{Z}exp\left[ln\pi_k + \sum_{j=1}^{D}(x_j(ln\theta_{jk} - ln(1-\theta_{jk})) + ln(1-\theta_{jk}))\right] \end{aligned} \tag{1}$$

   where $Z = P(x) = \sum_{k=0}^{1} P(y=k)P(x|y=k)$

3. **Answer**:

   $P(y=1|x) = \frac{1}{1+\frac{P(y=0,x)}{P(y=1,x)}} = \frac{1}{1+\frac{P(y=0|x)}{P(y=1|x)}}$

   $\frac{P(y=0|x)}{P(y=1|x)} = exp\left[ln(\frac{P(y=0|x)}{P(y=1|x)})\right] = exp\left[lnP(y=0|x) - lnP(y=1|x)\right]$

   $lnP(y=0|x) - lnP(y=1|x)$

   $= ln(1-\pi) + \sum_{j=1}^{D}(x_j(ln\theta_{j0} - ln(1-\theta_{j0})) + ln(1-\theta_{j0})) - \left[ln\pi + \sum_{j=1}^{D}(x_j(ln\theta_{j1} - ln(1-\theta_{j1})) + ln(1-\theta_{j1}))\right]$

   $= ln\frac{1-\pi}{\pi} + \sum_{j=1}^{D}(x_j ln\frac{\theta_{j0}(1-\theta_{j1})}{\theta_{j1}(1-\theta_{j0})} + ln\frac{1-\theta_{j0}}{1-\theta_{j1}})$

   $= ln\frac{1-\pi}{\pi} + \sum_{j=1}^{D} ln\frac{1-\theta_{j0}}{1-\theta_{j1}} + ln\frac{\boldsymbol{\theta_0(1-\theta_1)}}{\boldsymbol{\theta_1(1-\theta_0)}}\mathbf{x}$

   So we have

   $w_0 = -ln\frac{1-\pi}{\pi} - \sum_{j=1}^{D} ln\frac{1-\theta_{j0}}{1-\theta_{j1}}$

$$\mathbf{w} = (ln\frac{\theta_{10}(1-\theta_{11})}{\theta_{11}(1-\theta_{10})}, ln\frac{\theta_{20}(1-\theta_{21})}{\theta_{21}(1-\theta_{20})}, ..., ln\frac{\theta_{D0}(1-\theta_{D1})}{\theta_{D1}(1-\theta_{D0})})^T$$
$$\mathbf{x} = (x_1, x_2, ..., x_D)^T$$