# Introduction to Data Analytics

Xin Gao
Xin.gao@kaust.edu.sa
July 29, 2022
SDU

# An Example

- Let's learn classifiers by learning P(Y|X)
- Y = grades, X = <attend lectures, work hard>

| Attend Lectures | Work Hard | P(A\|AL, WH) | P(B\|AL, WH) |
|---|---|---|---|
| Y | Y | 0.86 | 0.14 |
| Y | N | 0.72 | 0.28 |
| N | Y | 0.58 | 0.42 |
| N | N | 0.13 | 0.87 |

- We need 4 parameters because of sum-to-one rule

# An Example

- Suppose $X = <X_1, X_2, \ldots, X_n>$, where $X_i$ and Y are random variables

- To estimate P(Y|X), we'll need $2^n$ parameters, even if we use the sum-to-one rule

- If we have 30 $X_i$'s, we will need $2^{30} > 1$billion parameters!

# Reduce Parameters

- From Bayes rule, we know

$$P(Y|X) = P(X|Y)\,P(Y) \,/\, P(X)$$

- For the classification rule, we know

$$1 \gtreqless \frac{P(Y=A|X)}{P(Y=B|X)} = \frac{P(X|Y=A)P(Y=A)}{P(X|Y=B)P(Y=B)}$$

- Now, how many parameters do we need?
  - $2^{n+1}-2$ for conditional probability and 1 for prior, even more!!

# Bayes Rule

- $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$
- This is shorthand for

$$P(Y = y_i | X = x_j) = \frac{P(X = x_j | Y = y_i) P(Y = y_i)}{P(X)}$$

$$= \frac{P(X = x_j | Y = y_i) P(Y = y_i)}{\sum_k P(X = x_j | Y = y_k) P(Y = y_k)}$$

# Naive Bayes

- Naive Bayes assumes

$$P(X_1, X_2, \dots, X_n | Y) = \prod_i P(X_i | Y)$$

- That is, $X_i$ and $X_j$ are conditionally independent given Y, for all $i \neq j$

# Recall Conditional Independence

- Definition: X is conditionally independent of Y given Z, if the probability distribution governing X is independent of the value of Y, given the value of Z

$$P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)$$

- We can often write P(X|Y, Z) = P(X|Z)
  - E.g., P(Thunder|Rain, Lightning) = P(Thunder|Lightning)

# Naive Bayes

- Naive Bayes uses assumption that $X_i$ are conditionally independent given Y
- Thus, $P(X_1, X_2|Y) = P(X_1|Y)P(X_2|Y)$
- In general, $P(X_1, X_2, \ldots, X_n|Y) = \prod_i P(X_n|Y)$

- Now, how many parameters do we need?
  - 2n!!
  - Without conditional independence, $2^{n+1}-2$

# Naive Bayes

- Bayes rule:

$$P(Y = y_i | X_1, X_2, \ldots, X_n) = \frac{P(X_1, X_2, \ldots, X_n | Y = y_i) P(Y = y_i)}{\sum_k P(X_1, X_2, \ldots, X_n | Y = y_k) P(Y = y_k)}$$

- By conditional independence:

$$P(Y = y_i | X_1, X_2, \ldots, X_n) = \frac{P(Y = y_i) \prod_j P(X_j | Y = y_i)}{\sum_k P(Y = y_k) \prod_j P(X_j | Y = y_k)}$$

- The classification rule is:

$$Y^{new} \leftarrow \operatorname{argmax}_{y_i} P(Y = y_i) \prod_j P(X_j^{new} | Y = y_i)$$

# Naive Bayes

- Define two parameters:
  - For each value $y_k$

$$\pi_k = P(Y = y_k)$$

  - For each value $x_{ij}$ of each attribute $X_i$

$$\theta_{ijk} = P(X_i = x_{ij} | Y = y_k)$$

- The classification rule is:

$$Y^{new} \leftarrow \mathrm{argmax}_{y_k} \pi_k \prod_i \theta_{ijk}$$

# Naive Bayes – Discrete X and Y

- Maximum likelihood estimation (MLE)
  - $\widehat{\pi_k} = P(Y = y_k) = \frac{\#D\{Y=y_k\}}{|D|}$
  - $\widehat{\theta_{ijk}} = P\left(X_i = x_{ij} | Y = y_k\right) = \frac{\#D\{X_i=x_{ij} \wedge Y=y_k\}}{\#D\{Y=y_k\}}$

# Example

- A new student? $P(M|P, Q, K)$
  - $M = 1$ iff a new student          $P = 1$ iff working on DM/ML
  - $Q = 1$ iff prepare for qual          $K = 1$ iff a Saudi citizen
- How many parameters do we need to estimate?

# Example

- A new student? P(M|P, Q, K)
  - M = 1 iff a new student — P = 1 iff working on DM/ML
  - Q = 1 iff prepare for qual — K = 1 iff a Saudi citizen
- How many parameters do we need to estimate?

P(M=1): P(M=0):

P(P=1|M=1): P(P=0|M=1):

P(P=1|M=0): P(P=0|M=0):

P(Q=1|M=1): P(Q=0|M=1):

P(Q=1|M=0): P(Q=0|M=0):

P(K=1|M=1): P(K=0|M=1):

P(K=1|M=0): P(K=0|M=0):

$$\frac{P(M=1)P(P=1|M=1)P(Q=1|M=1)P(K=1|M=1)}{P(M=0)P(P=1|M=0)P(Q=1|M=0)P(K=1|M=0)}$$

# An Issue

- If unlucky, our MLE for $P(X_i|Y)$ might be zero
  - One zero causes the entire probability to be zero!
  - Overfitting!

- How to avoid such issue?
  - Maximum a posteriori (MAP) estimation

# MAP for Naive Bayes

- MLE
  - $\widehat{\pi_k} = P(Y = y_k) = \frac{\#D\{Y=y_k\}}{|D|}$
  - $\widehat{\theta_{ijk}} = P(X_i = x_{ij} | Y = y_k) = \frac{\#D\{X_i=x_{ij} \wedge Y=y_k\}}{\#D\{Y=y_k\}}$
- MAP
  - $\widehat{\pi_k} = P(Y = y_k) = \frac{\#D\{Y=y_k\}+\alpha_k}{|D|+\sum_m \alpha_m}$
  - $\widehat{\theta_{ijk}} = P(X_i = x_{ij} | Y = y_k) = \frac{\#D\{X_i=x_{ij} \wedge Y=y_k\}+\alpha'_k}{\#D\{Y=y_k\}+\sum_m \alpha'_m}$

**Only difference: imaginary examples**

# MAP for Naive Bayes

- How to get $\alpha$'s?
  - $\alpha$'s are just priors of the parameters
  - Recall from Beta distribution…

- If N is big enough, prior is "forgotten"
- If N is small, prior is important

# Another Issue

- $X_i$'s are usually not really conditionally independent
  - We still use naive Bayes in many cases, and it usually works well
    - Often the right classification, even when not the right probability
  - What is the effect on estimated $P(Y|X)$?
    - Special case, what if we only have two $X_i$'s, and $X_1=X_2$?