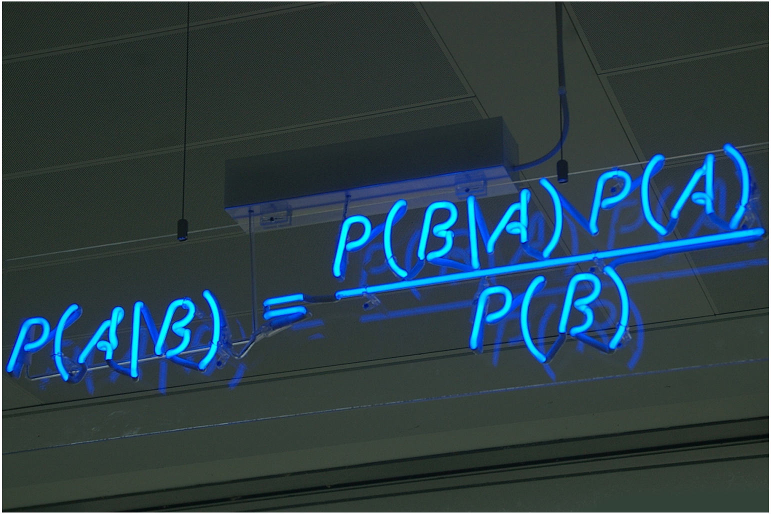
# CPSC 8420 Advanced Machine Learning
# Week 10: Bayesian Decision Theory

Dr. Kai Liu

October 22, 2020

# Naive Bayes



# Naive Bayes

When event $X, Y$ are independant, then:

$$P(X, Y) = P(X)P(Y). \tag{1}$$

Now let's take a look at the definition of Conditional Probability:

$$\begin{aligned} P(Y|X) &= P(X, Y)/P(X), \\ P(X|Y) &= P(X, Y)/P(Y). \end{aligned} \tag{2}$$

Based on which we have
$P(X, Y) = P(Y|X) * P(X) = P(X|Y) * P(Y)$, which implies:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{3}$$

# Naive Bayes

$$P(X) = \sum_k P(X|Y = Y_k)P(Y_k), \quad s.t. \sum_k P(Y_k) = 1 \tag{4}$$

therefore, we have:

$$P(Y_k|X) = \frac{P(X|Y_k)P(Y_k)}{\sum_k P(X|Y = Y_k)P(Y_k)} \tag{5}$$

# Who is Bayes

Thomas Bayes (1701? – 1761), he was elected as a Fellow of the Royal Society in 1742. His work was popularised by Laplace.

# Frequentist vs. Bayesian

(a) Karl Pearson
(27 March 1857 – 27 April 1936)

(b) Sir Ronald Aylmer Fisher
(17 February 1890 – 29 July 1962)

# Naive Bayes

$$P(Y_k|X) = \frac{P(X|Y_k)P(Y_k)}{\sum_k P(X|Y = Y_k)P(Y_k)}$$

We have two colored (red and blue) boxes, the red one has 2 apples and 6 oranges while the blue box contains one orange and 3 apples. Assume the probability to take from the red box is 0.4:

- What is the overall probability that the selection procedure will pick an apple?
- Given that we have chosen an orange, what is the probability that the box we chose was the red one?

# Bayesian Model

we are given data:
$(x_1^{(1)}, x_2^{(1)}, ...x_n^{(1)}, y_1), (x_1^{(2)}, x_2^{(2)}, ...x_n^{(2)}, y_2), ...(x_1^{(m)}, x_2^{(m)}, ...x_n^{(m)}, y_m)$,
that is we have $m$ samples with $n$ features, also assume there are $K$ classes defined as $C_1, C_2, \ldots, C_k$.
From the training samples we can get $P(Y = C_k)(k = 1, 2, ...K)$, if we know conditional probability distribution
$P(X = x|Y = C_k) = P(X_1 = x_1, X_2 = x_2, ...X_n = x_n|Y = C_k)$, then we will know joint probability distribution $P(X, Y)$:

$$\begin{aligned} P(X, Y = C_k) &= P(Y = C_k)P(X = x|Y = C_k) \\ &= P(Y = C_k)P(X_1 = x_1, X_2 = x_2, ...X_n = x_n|Y = C_k) \end{aligned} \tag{6}$$

# Naive Bayes

To compute $P(X_1 = x_1, X_2 = x_2, ...X_n = x_n|Y = C_k)$ is difficult, as it is of dimension $n$ density distribution. Bayes theorem make an assumption that each feature is independent to each other, in addtion, the order doesn't matter. Therefore:

$$\begin{aligned} &P(X_1 = x_1, X_2 = x_2, ...X_n = x_n|Y = C_k) \\ =&P(X_1 = x_1|Y = C_k)P(X_2 = x_2|Y = C_k)...P(X_n = x_n|Y = C_k) \end{aligned} \tag{7}$$

Though the (Naive) presumption doesn't have to be 100% true, but it is not that wrong. Moreover, it makes the model very concise and computationally friendly. That's the reason we call it as **Naive Bayes**.

# How to Classify New Data?

How can we classify a new data $(x_1^{(test)}, x_2^{(test)}, ...x_n^{(test)})$?

It is equivalent to maximize $P(Y = C_k|X = X^{(test)})$, which we can make use of Bayes Theorem:

$$C_{result} = \underbrace{argmax}_{C_k} P(Y = C_k|X = X^{(test)}) \tag{8}$$

$$= \underbrace{argmax}_{C_k} P(X = X^{(test)}|Y = C_k)P(Y = C_k) \Big/ P(X = X^{(test)}) \tag{9}$$

# How to Classify?

For all classes to compute $P(Y = C_k | X = X^{(test)})$, the denominator is all the same, thus we can simplify the above equation as:

$$
\begin{aligned}
C_{result} &= \underbrace{argmax}_{C_k} P(X = X^{(test)} | Y = C_k) P(Y = C_k) \\
&= \underbrace{argmax}_{C_k} P(Y = C_k) \prod_{j=1}^{n} P(X_j = X_j^{(test)} | Y = C_k)
\end{aligned} \tag{10}
$$

where $P(Y = C_k) = \frac{|D_k|}{|D|}$. The only problem remains how to compute $P(X_j = X_j^{(test)} | Y = C_k)$.

# Computing Conditional Probability Distribution

We discuss two cases for $P(X_j = X_j^{(test)} | Y = C_k)$:

- If the data is discrete, then $P(X_j = X_j^{(test)} | Y = C_k) = \frac{|D_{k,j}|}{|D_k|}$
- If the data is continous and assume it is Normal Distribution with $\mu_k, \sigma_k^2$ being the expecation and variance respectively. Then $P(X_j = X_j^{(test)} | Y = C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} exp\left( - \frac{(X_j^{(test)} - \mu_k)^2}{2\sigma_k^2} \right)$.

# Example

Given 5 cases whether go hiking or not:

- **P**: sunny, unhappy, free
- **P**: sunny, happy, free
- **N**: rainy, happy, free
- **P**: sunny, unhappy, busy
- **N**: rainy, unhappy, busy

please determine whether go hiking or not if given: "windy, happy, free" by making use of *Naive Bayes* rule.

# Laplacian Smoothing

Sometimes, some features may not appear in traning samples, which may lead $P(X_j = X_j^{(test)} | Y = C_k) = 0$, to avoid that, Laplacian Smoothing is introduced:

$$P(X_j = X_j^{(test)} | Y = C_k) = \frac{m_{kj^{test}} + \lambda}{m_k + O_j \lambda} \tag{11}$$

where $\lambda > 0$, and usually taken as 1. $O_j$ is the number of options for $X_j$.

# Example

Given 5 cases whether go hiking or not:

- **P**: sunny, unhappy, free
- **P**: sunny, happy, free
- **N**: rainy, happy, free
- **P**: sunny, unhappy, busy
- **N**: rainy, unhappy, busy

please determine whether go hiking or not if given: "windy, happy, free" by making use of *Naive Bayes* rule.

# Bayes Decision Boundary

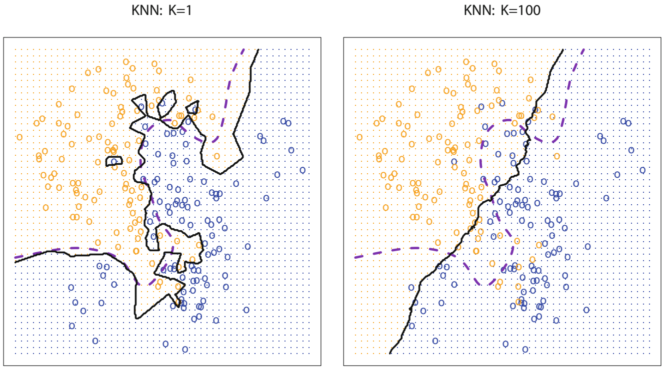



**FIGURE 2.16.** *A comparison of the KNN decision boundaries (solid black curves) obtained using $K = 1$ and $K = 100$ on the data from Figure 2.13. With $K = 1$, the decision boundary is overly flexible, while with $K = 100$ it is not sufficiently flexible. The Bayes decision boundary is shown as a purple dashed line.*

How can we determine Bayes Decision Boundary?

# Bayes Decision Boundary

$$\frac{P(C_1 \mid \mathrm{X})}{P(C_2 \mid \mathrm{X})} = \frac{P(\mathrm{X} \mid C_1)}{P(\mathrm{X} \mid C_2)} \frac{P(C_1)}{P(C_2)} \tag{12}$$

- First, compute $\mu_k, \sigma_k^2$ for training data points.
- Then, compute prior probabilty $P(Y = C_k) = m_k/m$.
- $P(X_j = x_j | Y = C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} exp\left( - \frac{(x_j - \mu_k)^2}{2\sigma_k^2} \right)$

Given data and lables:
$X = ([-1, -1], [-2, -1], [-3, -2], [1, 1], [2, 1], [3, 2])$, $Y = ([1, 1, 1, 2, 2, 2])$, how to decide the class of new data $[-0.8, -1]$?

# Semi-naive Bayes Classifiers

Naive Bayes assumes that each feature is independent, however, this is somewhat unrealistic. For example: word 'I' and 'am' are closely related. So the independent property can be relaxed such as: One-Dependent Estimator (ODE), which assumes each attribute is dependent with at most another parent attribute, that is:

$$P(c|x) \propto P(c) \prod_{j=1}^{d} P(x_j|c, pa_j) \tag{13}$$