

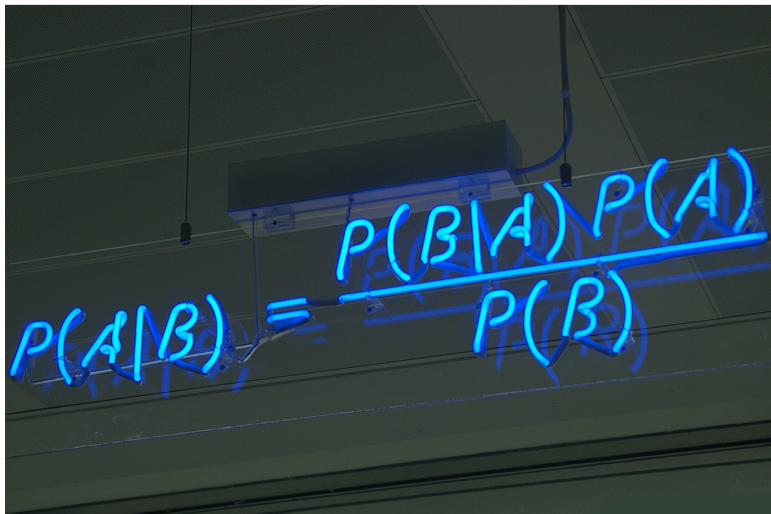
CPSC 8420 Advanced Machine Learning

Week 10: Bayesian Decision Theory

Dr. Kai Liu

October 22, 2020

Naive Bayes



A photograph of a blue neon sign mounted on a dark ceiling. The sign displays the Naive Bayes formula in a stylized, glowing blue font. The formula is written as $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$. The sign is slightly tilted and has some faint, illegible markings on the ceiling behind it.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Naive Bayes

When event X, Y are independent, then:

$$P(X, Y) = P(X)P(Y). \quad (1)$$

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

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

Based on which we have

$P(X, Y) = P(Y|X) * P(X) = P(X|Y) * P(Y)$, which implies:

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

Naive Bayes

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

therefore, we have:

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

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:

$(x_1^{(1)}, x_2^{(1)}, \dots, x_n^{(1)}, y_1), (x_1^{(2)}, x_2^{(2)}, \dots, x_n^{(2)}, y_2), \dots, (x_1^{(m)}, x_2^{(m)}, \dots, x_n^{(m)}, y_m)$,
that is we have m samples with n features, also assume there are K classes defined as C_1, C_2, \dots, C_k .

From the training samples we can get $P(Y = C_k) (k = 1, 2, \dots, K)$, if we know conditional probability distribution

$P(X = x|Y = C_k) = P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n|Y = C_k)$, then we will know joint probability distribution $P(X, Y)$:

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

Naive Bayes

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

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

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 $(x_1^{(test)}, x_2^{(test)}, \dots, x_n^{(test)})$?

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

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

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

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:

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

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 continuous and assume it is Normal Distribution with μ_k, σ_k^2 being the expectation 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 comments with positive/negative labels:

- **P**: extremely impressive
- **N**: extremely boring
- **P**: boring but impressive
- **N**: impressive but boring
- **P**: impressive and amazing

please determine the sentiment of a new testing comment: “amazing and impressive but somewhat boring” by making use of *Naive Bayes* rule.

Laplacian Smoothing

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

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

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

Example

Given 5 comments with positive/negative labels:

- **P**: extremely impressive
- **N**: extremely boring
- **P**: boring but impressive
- **N**: impressive but boring
- **P**: impressive and amazing

please determine the sentiment of a new testing comment: “amazing and impressive but somewhat boring” by making use of *Naive Bayes* rule.

Bayes Decision Boundary

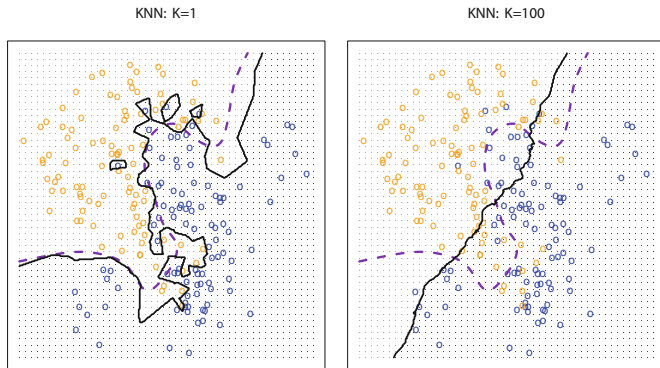


FIGURE 2.16. A comparison of the KNN decision boundaries (solid black curves) obtained using $K = 1$ and $K = 100$ on the data from Figure 2.13. With $K = 1$, the decision boundary is overly flexible, while with $K = 100$ it is not sufficiently flexible. The Bayes decision boundary is shown as a purple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

How can we determine Bayes Decision Boundary?

Bayes Decision Boundary

$$\frac{P(C_1 | X)}{P(C_2 | X)} = \frac{P(X | C_1) P(C_1)}{P(X | C_2) P(C_2)} \quad (12)$$

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

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:

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

