CS 4850/6850: Introduction to Machine Learning **Fall 2018**

Topic 6: Naive Bayes

INSTRUCTOR: DANIEL L. PIMENTEL-ALARCÓN

6.1 Introduction

Naive Bayes is one of the simplest classification methods. The main idea is to choose the class with the highest *posterior* probability (which accounts for the *prior* probability of each class, and the *conditional* probability of each class given each sample). Naive Bayes makes the strong assumption that features are independent, which is rarely true in practice. However, it tends to work well regardless.

6.2 Bayes Rule

Definition 6.1 (Conditional probability)**.** Let A, B be two events. The *conditional probability* that A occurs given B occurred is

$$\mathbb{P}(A|B) \;:=\; \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Example 6.1. Consider a 6-faced fair die. Let $A = \{1, 2\}$ be the event that the die rolls either 1 or 2, and similarly for $B = \{2, 3\}$. The probability that A occurs is $\mathbb{P}(A) = 1/3$. However, if you already know that B occurred, then the conditional probability that A also ocurred increases to

$$\mathbb{P}(A|B) \;=\; \frac{1/6}{1/3} \;=\; \frac{1}{2}.$$

Given the conditional probability $\mathbb{P}(A|B)$, Bayes rule gives us a formula for the *posterior* probability, $\mathbb{P}(B|A)$.

Definition 6.2 (Bayes rule)**.** Let A, B be two events. Then

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

Bayes rule plays a crucial role in modern applications.

Example 6.2. Geneticists have determined that 90% of the people with disease B have gene A active, i.e., $\mathbb{P}(A|B) = 0.9$. If you sequence your genome and find out that your gene A is active, what is the probability that you develop disease B? In other words, what is $\mathbb{P}(B|A)$? At first glance you might think it is very likely that you will develop disease B. However, to determine this you need to know $\mathbb{P}(A)$ and $\mathbb{P}(B)$. Of the whole population, if only 5% have disease B, while 45% have gene A active, what is $\mathbb{P}(B|A)$? This is a simple application of Bayes rule:

$$\mathbb{P}(B|A) \;=\; \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)} \;=\; \frac{(0.9)(0.05)}{0.45} \;=\; 0.1$$

Definition 6.3 (Independent events)**.** Let A, B be two events. We say A and B are *independent* if

$$\mathbb{P}(A|B) = \mathbb{P}(A).$$

Example 6.3. Consider two fair dice. Let A be the event that the first die is 1; let B be the event that the second die is 1. Then

$$\mathbb{P}(A|B) \;=\; \frac{\mathbb{P}(A=1 \cap B=1)}{\mathbb{P}(B=1)} = \frac{1/36}{1/6} = \frac{1}{6} = \mathbb{P}(A).$$

Hence the events A and B are independent. This matches our intuition that one die has no influence on the outcome of the other.

6.3 Naive Bayes

In naive Bayes one has a collection of N training pairs $\{(\mathbf{x}_i, k_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^d$ contains the d features of the i^{th} sample (e.g., glucose level, height, gender, etc.), and $k_i \in \{1, \ldots, K\} =: [K]$ denotes the class to which sample i belongs.

Given a new sample $\mathbf{x}$, the goal is to determine to which class it belongs. The main idea is to choose the class with the largest posterior probability, i.e.,

$$k^\star \;=\; \arg\max_{k \in [K]} \; \mathbb{P}(k|\mathbf{x}).$$

Using Bayes rule we have that:

$$k^\star \;=\; \arg\max_{k \in [K]} \; \frac{\mathbb{P}(\mathbf{x}|k)\mathbb{P}(k)}{\mathbb{P}(\mathbf{x})} \;=\; \arg\max_{k \in [K]} \; \mathbb{P}(\mathbf{x}|k)\mathbb{P}(k),$$

where the last step follows because $\mathbb{P}(\mathbf{x})$ does not depend on k. Let x_j denote the j^{th} feature of $\mathbf{x}$, with $j = 1, \ldots, d$. Naive Bayes assumes that the x_j's are independent. This implies that

$$\mathbb{P}(\mathbf{x}|k) \;:=\; \mathbb{P}(x_1, x_2, \ldots, x_d|k) \;=\; \prod_{j=1}^{d} \mathbb{P}(x_j|k).$$

Consequently,

$$k^\star \;=\; \arg\max_{k \in [K]} \; \mathbb{P}(k) \prod_{j=1}^{d} \mathbb{P}(x_j|k),$$

With this expression, it all boils down to estimating $\mathbb{P}(\mathrm{k})$ and $\mathbb{P}(\mathrm{x_j}|\mathrm{k})$ for every k, which we can do using training data.

6.4 Example

Consider the following dataset:

Bacteria	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Gene 1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1
Gene 2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Gene 3	0	1	1	0	0	0	1	1	1	1	0	0	0	1	0	1
Gene 4	1	1	0	0	1	1	1	0	1	1	0	0	1	0	0	0
Gene 5	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0
Gene 6	0	0	1	0	0	1	1	1	0	0	1	0	1	1	0	0
Gene 7	0	0	0	1	0	0	0	1	1	0	0	0	1	0	1	0
Gene 8	0	1	0	0	0	0	1	1	1	0	1	1	0	0	0	1
Gene 9	0	0	1	0	0	1	0	1	0	0	1	0	1	1	1	1
Gene 10	0	1	0	1	1	0	0	1	1	0	0	1	0	1	0	1
Class	**1**	**3**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**3**	**1**

We can estimate $\mathbb{P}(\mathrm{k})$ as the fraction of samples that fall under each class:

$$
\begin{aligned}
\mathbb{P}(\mathrm{k}=1) &= \frac{12}{16} = \frac{6}{8}, \\
\mathbb{P}(\mathrm{k}=2) &= \frac{2}{16} = \frac{1}{8}, \\
\mathbb{P}(\mathrm{k}=3) &= \frac{2}{16} = \frac{1}{8}.
\end{aligned}
$$

Similarly, we can model each gene $\mathrm{x_j}$ as a Bernoulli(p_j) random variable, and compute $p_\mathrm{j} := \mathbb{P}(\mathrm{x_j}=1|\mathrm{k})$ as the fraction of 1s in each class, to obtain the following conditional probability matrix:

$$
\mathbb{P}(\mathrm{x_j}=1|\mathrm{k}) \;=\; \begin{array}{c c}
 & \begin{array}{ccc} 1 & 2 & 3 \end{array} \\
\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \end{array} &
\begin{bmatrix}
{}^{11}\!/_{12} & 1 & 1 \\
{}^{11}\!/_{12} & 0 & 0 \\
{}^{7}\!/_{12} & 0 & {}^{1}\!/_{2} \\
{}^{1}\!/_{2} & {}^{1}\!/_{2} & {}^{1}\!/_{2} \\
{}^{1}\!/_{12} & {}^{1}\!/_{2} & {}^{1}\!/_{2} \\
{}^{7}\!/_{12} & 0 & 0 \\
{}^{1}\!/_{3} & 0 & {}^{1}\!/_{2} \\
{}^{5}\!/_{12} & {}^{1}\!/_{2} & {}^{1}\!/_{2} \\
{}^{7}\!/_{12} & 0 & {}^{1}\!/_{2} \\
{}^{5}\!/_{12} & 1 & {}^{1}\!/_{2}
\end{bmatrix}
\end{array}.
$$

Given a new sample:

$$\mathbf{x} = \begin{array}{c c} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} & \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \end{array} \end{array},$$

all we have to do is compute the posterior probability of each class, given this sample:

$$\begin{aligned}
\mathbb{P}(\mathrm{k}=1|\mathbf{x}) &\propto \mathbb{P}(\mathrm{k}=1)\prod_{\mathrm{j}=1}^{\mathrm{d}}\mathbb{P}(\mathrm{x_j}|\mathrm{k}) \\
&= \frac{6}{8}\cdot\mathbb{P}(\mathrm{x}_1=1|\mathrm{k}=1)\cdot\mathbb{P}(\mathrm{x}_2=0|\mathrm{k}=1)\cdot\mathbb{P}(\mathrm{x}_3=0|\mathrm{k}=1) \\
&\qquad \mathbb{P}(\mathrm{x}_4=1|\mathrm{k}=1)\cdot\mathbb{P}(\mathrm{x}_5=0|\mathrm{k}=1)\cdot\mathbb{P}(\mathrm{x}_6=1|\mathrm{k}=1) \\
&\qquad \mathbb{P}(\mathrm{x}_7=1|\mathrm{k}=1)\cdot\mathbb{P}(\mathrm{x}_8=0|\mathrm{k}=1)\cdot\mathbb{P}(\mathrm{x}_9=1|\mathrm{k}=1) \\
&\qquad \mathbb{P}(\mathrm{x}_{10}=1|\mathrm{k}=1) \\
&= \frac{6}{8}\cdot\frac{11}{12}\cdot\frac{1}{12}\cdot\frac{5}{12}\cdot\frac{1}{2}\cdot\frac{11}{12}\cdot\frac{7}{12}\cdot\frac{1}{3}\cdot\frac{7}{12}\cdot\frac{7}{12}\cdot\frac{5}{12} \\
&= 3.0163\times 10^{-4}.
\end{aligned}$$

Similarly, we can compute $\mathbb{P}(\mathrm{k}=2|\mathbf{x})$ and $\mathbb{P}(\mathrm{k}=3|\mathbf{x})$ (what values do you obtain?), and choose the k with the largest posterior probability (how would you classify this new sample $\mathbf{x}$?).