Machine Learning (IV): Conditional Independence: Naive Bayes Classifiers and Other Examples

Le Wang

Conditional Distribution and Independence

$$p(x_i, y_j) = p(x_i)p(y_j)$$

Conditional Distribution and Independence

$$
\begin{aligned}
p(x_i, y_j) &= p(x_i)p(y_j) \\
\frac{p(x_i, y_j)}{p(x_i)} &= p(y_j)
\end{aligned}
$$

Conditional Distribution and Independence

$$p(x_i, y_j) = p(x_i)p(y_j)$$
$$\frac{p(x_i, y_j)}{p(x_i)} = p(y_j)$$

In other words

$$p(y_j \mid x_i) = p(y_j)$$

Independence implies that conditional distribution is marginal distribution!

Intuitively, no predictive power at all! as it should be for independent variables!

Unconditional Independence to Conditional Independence

To identify the causal effects of x on y, we will later rely on variants of the assumption of **conditional independence** (it is not the same as **unconditional independence**).

$$p(x, y|z) = p(x|z)p(y|z) \quad \text{for all } x, y, z$$
$$p(y|x, z) = p(y|z)$$

Interpretation: The conditional distribution of Y, given X, Z is in fact completely determined by the value of Z alone, Y being superfluous once Z is given.

Visualization of Conditional Independence

Question: Can you think of a way to visualize the conditional independence for all possible combinations?

Visualization of Conditional Independence

Question: Can you think of a way to visualize the conditional independence for all possible combinations?

A scatter plot of conditional probability and the product of two marginal probabilities with a 45 degree line. Why?

Further Applications

As we will see later, conditional independence plays an important role in causal inference, designing different experiments (e.g., conditional randomization), and resolving many of the empirical paradoxes (e.g., Simpson's paradox).

Conditional independence is, sometimes, a direct implication of economic theory. For example, in the literature of insurance, the presence of positive conditional dependence between coverage and risk is known to be a direct consequence of adverse selection under information asymmetry [e.g., Chiappori and Salanie (2000)].

The literature of testing conditional independence for continuous variables appears rather recent and includes relatively few researches [Song, 2009, Testing Conditional Independence via Transforms, Annals of Statistics, Vol. 37, No. 6B, 4011–4045.]

Application (I): The Problem of Operating a "Fair" Admission Policy

An important issue: Operating a "fair" procedure for the selection of minority group members for universeity admission. One solution is to require that the probability of such selection should depend only on the academic promise of the candidates, and not on race, sex and so on. How to examine whether this criterion is indeed met?

Let $Y = 1$ (selection) $Y = 0$ (rejection), let X denote sex or race, and let Z be a test-score or other measures of academic promise. Our goal is then to test

$$Y \perp X \mid Z$$

Given Z (academic peformance), there should not be any predictive power of X (sex or race) for Y (admission or not).

Example: `mv05_cond_indep01.R`

Application (II): Markov Chain Assumption

Recall that a **Markov Chain** is a stochastic process for whic, given the current state, future states of the random variable Y are independent of past states.

$$\Pr[Y_{t+1} = y \mid y_t, y_{t-1}, \ldots] = \Pr[Y_{t+1} = y \mid y_t]$$

Applications

Combining **conditional independence** and **Bayes' Rule**

1. **Statistical Language Model**
2. Naive Bayes Classifier
3. Predicting Unobservable Ethnicity in Political Science.

Application (III): Statistical Language Model

$$\text{Sender} \stackrel{s_1, s_2, s_3, \dots}{\Longrightarrow} \text{Chanel} \stackrel{o_1, o_2, o_3, \dots}{\Longrightarrow} \text{Receiver}$$

You send some signals to the receiver. The task of communication is to decode whatever signals the receiver receives. In other words, you would like to back out the original meaning $(s_1, s_2, s_3, \dots)$ of the received cell-phone signals $(o_1, o_2, o_3, \dots)$.

Nearly all natural language processing problems can be thought of as decoding the communication process.

This model is called **Acoustic Model** in voice recognition, **Translation Model** in machine learning, **Correction Model** in autocorrection.

1. Speech recognition: You say something, and the computer receives the digits and cracks out what you say.
2. Translations between languages

How is this related to what we have learned? Let me translate for you.

Just look for the most likely $s_1, s_2, s_3 \ldots$!

$$\Pr[s_1, s_2, s_3, \ldots | o_1, o_2, o_3, \ldots]$$

With **Conditional Independence**, **Bayes' Rule** and **Hidden Markov Model**, we can solve this.

Lets type in Goolge search “bomer tooner” (o_1, o_2). What would happen?

$$\Pr[s_1, s_2, s_3, \ldots | o_1, o_2, o_3, \ldots] =$$

$$\Pr[s_1, s_2, s_3, \ldots | o_1, o_2, o_3, \ldots] = \frac{\Pr[o_1, o_2, o_3, \ldots | s_1, s_2, s_3, \ldots] \cdot \Pr[s_1, s_2, s_3, \ldots]}{\Pr[o_1, o_2, o_3, \ldots]}$$

When I recevie the signal, $o_1, o_2, o_3 \ldots$ is already known, we can then ignore it. Because it is a constant, our goal is to maximize the quantity above. Then, the question becomes maximizing the following

$$\Pr[o_1, o_2, o_3, \ldots | s_1, s_2, s_3, \ldots] \cdot \Pr[s_1, s_2, s_3, \ldots]$$

How can we solve this?

Solution: Hidden Markov Model

Hidden Markov Model:

Leonard E. Baum and others developed Hidden Markov Model in 60s and 70s. The idea is simple:

Hidden Markov Model:

Leonard E. Baum and others developed Hidden Markov Model in 60s and 70s. The idea is simple:

We do not observe the original messages (hence **Hidden**), but I assume that it follows a Markov Model.

Hidden Markov Model:

Leonard E. Baum and others developed Hidden Markov Model in 60s and 70s. The idea is simple:

We do not observe the original messages (hence **Hidden**), but I assume that it follows a Markov Model.

In the meantime, even though I do not observe the original message, I assume that for every message s_t, I could observe a signal o_t. This signal o_t depends on s_t and **only** s_t!

$$\Pr[o_1, o_2, o_3, \ldots | s_1, s_2, s_3, \ldots] \cdot \Pr[s_1, s_2, s_3, \ldots]$$

1. Conditional Independence:

$$\Pr[o_1, o_2, o_3, \ldots | s_1, s_2, s_3, \ldots] = \Pr[o_1|s_1] \cdot \Pr[o_2|s_2] \cdots$$

2. (Hidden) Markov Model:

$$\Pr[s_1, s_2, s_3, \ldots] = \Pr[s_2|s_1] \cdot \Pr[s_3|s_2] \cdots$$

We can solve this problem using, e.g., Viterbi Algorithm

In the 70s, James and Janet Baker (co-founders of Dragon) propose using Hidden Markov model for voice recognition. The misclassfication rates reduced from 30% to 10%.

Kai-fu Lee used this type of model to develop Sphinx.

Applications

Combining **conditional independence** and **Bayes' Rule**

1. Statistical Language Model
2. **Naive Bayes Classifier**
3. Predicting Unobservable Ethnicity in Political Science.

Application (IV): Naive Bayes Classifier

Suppose that we now have K predictors $(X_1, X_2, \ldots, X_K)$ and the outcome has J classes $y_1, y_2, \ldots, y_J$.

Application (IV): Naive Bayes Classifier

Suppose that we now have K predictors $(X_1, X_2, \ldots, X_K)$ and the outcome has J classes $y_1, y_2, \ldots, y_J$.

Suppose that the predictors take on the following values

$$X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k$$

For prediction, we need to know the following values

$$\begin{gathered}
\Pr[Y = y_1 \mid X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k] \\
\Pr[Y = y_2 \mid X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k] \\
\Pr[Y = y_3 \mid X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k]
\end{gathered}$$

Question:

What is your prediction for the employment status of a guy named Le Wang, a 38-year-old man born in Guangdong but living in Vermont?

$$\begin{aligned}&\Pr[\text{Unemployment} = 1 \mid \text{Name} \ = \ \text{Le Wang}, \text{Age} \ = 38,\\ &\text{Gender} \ = \text{Male}, \ \text{Place of Birth} \ = \ \text{Guangdong} \ ,\\ &\text{Place of Residence} \ = \ \text{Vermont} \ ]\end{aligned}$$

Question:

What is your prediction for the employment status of a guy named Le Wang, a 38-year-old man born in Guangdong but living in Vermont?

$$\Pr[\text{Unemployment} = 1 \mid \text{Name} = \text{Le Wang}, \text{Age} = 38,$$
$$\text{Gender} = \text{Male}, \text{ Place of Birth} = \text{Guangdong} ,$$
$$\text{Place of Residence} = \text{Vermont}]$$

$$= \frac{\Pr[\text{Unemployment} = 1, \text{Name} = \text{Le Wang}, \text{Age} = 38, \ldots]}{\Pr[\text{Name} = \text{Le Wang}, \text{Age} = 38, \ldots]}$$

Naive Bayes Classififiers:

1. **Step 1.** We can simplify this using **Bayes' Rule**
2. **Step 2.** We can simplify this using **Conditional Independence Assumption**

Step 1. We can simplify this using **Bayes' Rule**

$$\Pr[Y = y_1 \mid X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k]$$

$$= \frac{\Pr[Y = y_1, X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k]}{\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k]}$$

Step 1. We can simplify this using **Bayes' Rule**

$$\Pr[Y = y_1 \mid X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k]$$

$$= \frac{\Pr[Y = y_1, X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k]}{\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k]}$$

$$= \frac{\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k \mid Y = y_1] \cdot \Pr[Y = y_1]}{\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k]}$$

Step 1. We can simplify this using **Bayes' Rule**

$$\Pr[Y = y_1 \mid X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k]$$

$$= \frac{\Pr[Y = y_1, X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k]}{\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k]}$$

$$= \frac{\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k \mid Y = y_1] \cdot \Pr[Y = y_1]}{\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k]}$$

$$= \frac{\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k \mid Y = y_1] \cdot \Pr[Y = y_1]}{\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k]}$$

Step 2. We can simplify this using **Conditional Independence Assumption**

$$\frac{\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k \mid Y = y_1] \cdot \Pr[Y = y_1]}{\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k]}$$

Step 2. We can simplify this using **Conditional Independence Assumption**

$$\frac{\Pr[X_1 = x_1, X_2 = x_2, \dots, X_K = x_k \mid Y = y_1] \cdot \Pr[Y = y_1]}{\Pr[X_1 = x_1, X_2 = x_2, \dots, X_K = x_k]}$$

Suppose conditional on the class, every predictor is independent of each other.

Step 2. We can simplify this using **Conditional Independence Assumption**

$$\frac{\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k \mid Y = y_1] \cdot \Pr[Y = y_1]}{\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k]}$$

Suppose conditional on the class, every predictor is independent of each other.

$$\begin{aligned} &\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k \mid Y = y_1,] \\ = &\Pr[X_1 = x_1 \mid Y = y_1] \cdot \Pr[X_2 = x_2 \mid Y = y_1] \cdot, \\ &\ldots, \cdot \Pr[X_K = x_k \mid Y = y_1] \end{aligned}$$

Our Example (with conditional independence)

$$\begin{aligned}&\Pr[\text{Name } = \text{ Le Wang}, \text{Age } = 38,\\ &\text{Gender } = \text{Male}, \text{ Place of Birth } = \text{ Guangdong },\\ &\text{Place of Residence } = \text{ Vermont } \mid \text{Unemployment} = 1]\end{aligned}$$

Our Example (with conditional independence)

$$\begin{aligned}&\Pr[\text{Name} \;=\; \text{Le Wang}, \text{Age} \;= 38,\\ &\text{Gender} \;= \text{Male},\; \text{Place of Birth} \;=\; \text{Guangdong}\;,\\ &\text{Place of Residence} \;=\; \text{Vermont} \;\mid \text{Unemployment} = 1]\end{aligned}$$

$$\begin{aligned}&\Pr[\text{Name} \;=\; \text{Le Wang} \mid \text{Unemployment} = 1]\\ &\cdot \Pr[\text{Age} \;= 38 \mid \text{Unemployment} = 1]\\ &\cdot \Pr[\;\text{Gender} \;= \text{Male} \mid \text{Unemployment} = 1]\\ &\cdot \Pr[\;\text{Place of Birth} \;=\; \text{Guangdong} \;\mid \text{Unemployment} = 1]\\ &\cdot \Pr[\;\text{Place of Residence} \;=\; \text{Vermont} \;\mid \text{Unemployment} = 1]\end{aligned}$$

Step 3. Apply conditional independence and law of total probability (as we did before!)

$$\frac{\Pi_{i=1}^{K} \Pr[X_i = x_i \mid Y = y_1] \cdot \Pr[Y = y_1]}{\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k]}$$

How to get

$$\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k]$$

Step 3. Apply conditional independence and law of total probability (as we did before!)

$$\frac{\Pi_{i=1}^{K} \Pr[X_i = x_i \mid Y = y_1] \cdot \Pr[Y = y_1]}{\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k]}$$

How to get

$$\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k]$$

$$\begin{aligned}
&\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k] \\
=&\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k \mid Y = y_1] \cdot \Pr[Y = y_1] + \\
&\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k \mid Y = y_2] \cdot \Pr[Y = y_2] + \\
&\vdots \\
&\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k \mid Y = y_J] \cdot \Pr[Y = y_J]
\end{aligned}$$

Step 3. Apply conditional independence and law of total probability (as we did before!)

$$\frac{\Pi_{i=1}^{K} \Pr[X_i = x_i \mid Y = y_1] \cdot \Pr[Y = y_1]}{\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k]}$$

How to get

$$\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k]$$

$$\begin{aligned}
&\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k] \\
=&\Pi_{i=1}^{K} \Pr[X_i = x_i \mid Y = y_1] \cdot \Pr[Y = y_1] + \\
&\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k \mid Y = y_2] \cdot \Pr[Y = y_2] + \\
&\vdots \\
&\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k \mid Y = y_J] \cdot \Pr[Y = y_J]
\end{aligned}$$

Step 3. Apply conditional independence and law of total probability (as we did before!)

$$\frac{\Pi_{i=1}^{K} \Pr[X_i = x_i \mid Y = y_1] \cdot \Pr[Y = y_1]}{\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k]}$$

How to get

$$\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k]$$

$$\begin{aligned}
&\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k] \\
=&\Pi_{i=1}^{K} \Pr[X_i = x_i \mid Y = y_1] \cdot \Pr[Y = y_1] + \\
&\Pi_{i=1}^{K} \Pr[X_i = x_i \mid Y = y_1] \cdot \Pr[Y = y_2] + \\
&\vdots \\
&\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k \mid Y = y_J] \cdot \Pr[Y = y_J]
\end{aligned}$$

Step 3. Apply conditional independence and law of total probability (as we did before!)

$$\frac{\Pi_{i=1}^{K} \Pr[X_i = x_i \mid Y = y_1] \cdot \Pr[Y = y_1]}{\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k]}$$

How to get

$$\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k]$$

$$\begin{aligned}
&\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k] \\
=&\Pi_{i=1}^{K} \Pr[X_i = x_i \mid Y = y_1] \cdot \Pr[Y = y_1] + \\
&\Pi_{i=1}^{K} \Pr[X_i = x_i \mid Y = y_1] \cdot \Pr[Y = y_2] + \\
&\vdots \\
&\Pi_{i=1}^{K} \Pr[X_i = x_i \mid Y = y_1] \cdot \Pr[Y = y_J]
\end{aligned}$$

Step 3. Apply conditional independence and law of total probability (as we did before!)

$$\frac{\prod_{i=1}^{K} \Pr[X_i = x_i \mid Y = y_1] \cdot \Pr[Y = y_1]}{\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k]}$$

$$= \frac{\prod_{i=1}^{K} \Pr[X_i = x_i \mid Y = y_1] \cdot \Pr[Y = y_1]}{\sum_{j=1}^{J} \prod_{i=1}^{K} \Pr[X_i = x_i \mid Y = y_j] \cdot \Pr[Y = y_j]}$$

Our Example (with conditional independence)

$$
\begin{aligned}
&(\Pr[\text{Name} \;=\; \text{Le Wang} \mid \text{Unemployment} = 1] \\
&\cdot \Pr[\text{Age} \;= 38 \mid \text{Unemployment} = 1] \\
&\cdot \Pr[\text{ Gender } = \text{Male} \mid \text{Unemployment} = 1] \\
&\cdot \Pr[\text{ Place of Birth } \;=\; \text{ Guangdong } \mid \text{Unemployment} = 1] \\
&\cdot \Pr[\text{ Place of Residence } \;=\; \text{ Vermont } \mid \text{Unemployment} = 1]) \\
&\times \Pr[\text{Unemployment} = 1] +
\end{aligned}
$$

Our Example (with conditional independence)

$$
\begin{aligned}
&(\Pr[\text{Name} = \text{Le Wang} \mid \text{Unemployment} = 1] \\
&\cdot \Pr[\text{Age} = 38 \mid \text{Unemployment} = 1] \\
&\cdot \Pr[\text{Gender} = \text{Male} \mid \text{Unemployment} = 1] \\
&\cdot \Pr[\text{Place of Birth} = \text{Guangdong} \mid \text{Unemployment} = 1] \\
&\cdot \Pr[\text{Place of Residence} = \text{Vermont} \mid \text{Unemployment} = 1]) \\
&\times \Pr[\text{Unemployment} = 1] + \\
&(\Pr[\text{Name} = \text{Le Wang} \mid \text{Unemployment} = 0] \\
&\cdot \Pr[\text{Age} = 38 \mid \text{Unemployment} = 0] \\
&\cdot \Pr[\text{Gender} = \text{Male} \mid \text{Unemployment} = 0] \\
&\cdot \Pr[\text{Place of Birth} = \text{Guangdong} \mid \text{Unemployment} = 0] \\
&\cdot \Pr[\text{Place of Residence} = \text{Vermont} \mid \text{Unemployment} = 0]) \\
&\times \Pr[\text{Unemployment} = 0]
\end{aligned}
$$

Example: `mv05_cond_indep01.R`

Applications

Combining **conditional independence** and **Bayes' Rule**

1. Statistical Language Model
2. Naive Bayes Classifier
3. **Predicting Unobservable Ethnicity in Political Science.**'

Frontier Research: Predicting Unobservable Ethnicity

Motivation:

In both political behavior research and voting rights litigation, turnout and vote choice for different racial groups are often inferred using aggregate election results and racial composition. These predictions are often inaccurate.

How to reduce aggregation bias by predicting individual-level ethnicity from voter registration records.

Frontier Research: Predicting Unobservable Ethnicity

Motivation:

In both political behavior research and voting rights litigation, turnout and vote choice for different racial groups are often inferred using aggregate election results and racial composition. These predictions are often inaccurate.

How to reduce aggregation bias by predicting individual-level ethnicity from voter registration records.

Imai and Khanna (2016), Political Analysis

1. **Conditional Probability.**
2. **Bayes's rule** to combine the Census Bureau's Surname List with various information from geocoded voter registration records.

To classify someone's race:

$$\Pr[\text{race}]$$

To classify someone's race:

$$\Pr[\text{race}]$$
$$\Pr[\text{race}|\text{surname}]$$

To classify someone's race:

$$\begin{aligned} &\Pr[\text{race}] \\ &\Pr[\text{race}|\text{surname}] \\ &\Pr[\text{race}|\text{surname}, \text{residence}] \end{aligned}$$

We will use the maximum to forecast. If no further information, everyone is predicted to be white!

Purpose of this exercise

To validate the accuracy of the predictions of individual race, we will use the sample of $10,000$ registered voters from Florida.

In Florida, voters are asked to self-report their race when registering.

Datasets

1. Florida Registered Voter Data
2. Census Data

Dataset (I): Florida Registered Voter Data

Variable	Description
`surname`	Surname
`county`	County id of voter's residence
`VTD`	voting district id of voter's residence
`age`	age
`gender`	gender: $m = male$ and $f = female$
`race`	self-reported race

Dataset (II): Census Data

Variable	Description
`surname`	Surname
`count`	number of individuals with a specific surname
`pctwhite`	percentage of non-Hispanic whites among those who have a specific surname
`pctblack`	percentage of non-Hispanic blacks among those who have a specific surname
`pctapi`	percentage of non-Hispanic Asians and Pacific Islanders among those who have a specific surname
`pcthispanic`	percentage of Hispanic origin among those who have a specific surname