Machine Learning

Yuh-Jye Lee

Lab of Data Science and Machine Intelligence
Dept. of Applied Math. at NCTU

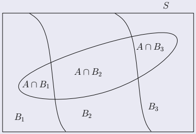
March 1, 2017

Bayes' Rule

Bayes' Rule

Assume that $\{B_1, B_2, \ldots, B_k\}$ is a partition of S such that $P(B_i) > 0$, for $i = 1, 2, \ldots, k$. Then

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)}.$$



Applying Baye's Rule to Classification

Credit Cards Scoring: Low-risk vs. High-risk

- According to the past transactions, some customers are low-risk in that they paid back their loan and the bank profited from them and other customers are high-risk in that they defaulted.
- We would like to *learn* the class "*high-risk customer*"
- We observe customer's *yearly income* and *savings*, which we represent by two *random variables* X_1 and X_2
- The *credibility of a customer* is denoted by a *Bernoulli* random variable C where $C = 1$ indicates a high-risk customer and $C = 0$ indicated a low-risk customer

Applying Baye's Rule to Classification

How to make the decision when a new application arrives?

- When a new application arrives with $X_1 = x_1$ and $X_2 = x_2$
- If we know the probability of C *conditioned on* the observation $X = [x_1,\ x_2]$ our decision will be
  - $C = 1$ if $P(C = 1|[x_1,\ x_2]) > 0.5$
  - $C = 0$ otherwise
- The probability of error we made based on this rule is

$$1 - \max\{P(C = 1|[x_1,\ x_2]), P(C = 0|[x_1,\ x_2])\} < 0.5$$

- Please note $P(C = 1|[x_1,\ x_2]) + P(C = 0|[x_1,\ x_2]) = 1$

The *Posterior Probability*: $P(C|\mathbf{x}) = \frac{P(C)P(\mathbf{x}|C)}{P(\mathbf{x})}$

- $P(C = 1)$ is called the *prior probability* that $C = 1$
- In our example, it corresponds to a probability that a customer is high-risk, *regardless* of the $\mathbf{x}$ value.
- It is called the *prior probability* because it is the knowledge we have *before* looking at the observation $\mathbf{x}$
- $P(\mathbf{x}|C)$ is called the *class likelihood* and is the *conditional probability* that an *event belonging to the class C* has the associated observation value $\mathbf{x}$
- $P(\mathbf{x})$, the *evidence* is the probability that an observation $\mathbf{x}$ to be seen, regardless of whether it is a positive or negative example

All above information can be extracted from the past transactions *(historical data)*

The *Posterior Probability*: $P(C|\mathbf{x}) = \frac{P(C)P(\mathbf{x}|C)}{P(\mathbf{x})}$

- Because of normalization by the evidence, the posteriors sum up to 1
- In our example, $P(X_1, X_2)$ is called the *joined probability* of two random variables X_1 and X_2
- Under the assumption, these two random variables X_1 and X_2 are *conditional probability independent*, we have $P(X_1, X_2|C) = P(X_1|C)P(X_2|C)$
- It is one of key assumptions of *Naive Bayes' Classifier*
- Although it is *over simplified* the problem it is very easy to use for real applications

Extend to Multi-class classification

- We have K mutually and exhaustive classes; $C_i, \; i = 1, 2, \ldots, K$
- For example, in *optical digit recognition*, the input is a *bitmap image* and there are 10 classes
- We can think of that these K classes define a *partition* of the *input space*
- Please refer to the slides of the *Partition Theorem* and *Baye's Rule*
- The Bayes' classifier choose the class with the highest posterior probability; that is we choose C_i if

$$P(C_i|\mathbf{x}) = \max_k P(C_k|\mathbf{x})$$

- Question: Is it very important to have $P(\mathbf{x})$, the evidence?

Naïve Bayes for Classification
Also Good for Multi-class Classification

- Estimate a *posteriori probability* of class label
- Let each *attribute* (variable) be a *random variable*. What is the probibility of

$$Pr(y = 1|\mathbf{x}) = Pr(y = 1|\mathbf{X}_1 = x_1, \mathbf{X}_2 = x_2, \ldots, \mathbf{X}_n = x_n)$$

- Naïve Bayes TWO not reasonable assumptions:
  - The importance of each attribute is *equal*
  - All attributes are *conditional probability independent* !

$$Pr(y = 1|\mathbf{x}) = \frac{1}{Pr(\mathbf{X} = \mathbf{x})} \prod_{i=1}^{n} Pr(\mathbf{X}_i = x_i|y = 1)$$

The Weather Data Example

Ian H. Witten & Eibe Frank, Data Mining

Outlook	Temperature	Humidity	Windy	Play(Label)
Sunny	Hot	High	False	-1
Sunny	Hot	High	True	-1
Overcast	Hot	High	False	+1
Rainy	Mild	High	False	+1
Rainy	Cool	Normal	False	+1
Rainy	Cool	Normal	True	-1
Overcast	Cool	Normal	True	+1
Sunny	Mild	High	False	-1
Sunny	Cool	Normal	False	+1
Rainy	Mild	Normal	False	+1
Sunny	Mild	Normal	True	+1
Overcast	Mild	High	True	+1
Overcast	Hot	Normal	False	+1
Rainy	Mild	High	True	-1

MLE for Bernoulli Distribution
play vs. not play

Likelihood Function

The probability to *observe* the random sample $\mathbf{X} = \{x^t\}_{t=1}^N$ is

$$\prod_{t=1}^{N} p^{x^t}(1-p)^{1-x^t}$$

Why don't we choose the parameter p which will maximize the probability for observing the random sample $\mathbf{X} = \{x^t\}_{t=1}^N$?

Based on MLE, we will choose the parameter p

$$p = \frac{\sum_{t=1}^N x^t}{N}$$

MLE for Multinomial Distribution

Multinomial Distribution: Sunny, Cloudy and Rainy

Consider the generalization of Bernoulli where instead of two possible outcomes, the outcome of a random event is one of k *classes*, each of which has a probability of occurring p_i and $\sum_{i=1}^{k} p_i = 1$. Let $x_1, x_2, \ldots, x_k$ be k indicator variables where $x_i = 1$ if the outcome is class i and $x_i = 0$ otherwise. *i.e.*,

$$P(x_1, x_2, \ldots, x_k) = \prod_{i=1}^{k} p_i^{x_i}$$

Let $\mathbf{X} = \{\mathbf{x}^\mathbf{t}\}_{t=1}^{N}$ be N independent radom experiments. Based on MLE, we will choose the parameter $\hat{p}_i$

$$\hat{p}_i = \frac{\sum_{t=1}^{N} x_i^t}{N}, \quad i = 1, 2, \ldots k$$

Probabilities for Weather Data Using Maximum Likelihood Estimation

Based on MLE, we will choose the parameter $\hat{p}_i$

$$\hat{p}_i = \frac{\sum_{t=1}^{N} x_i^t}{N}, \quad i = 1, 2, \ldots k$$

Outlook			Temp.			Humidity			Windy			Play	
Play	Yes	No		Yes	No		Yes	No		Yes	No	Yes	No
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	T	3/9	3/5	9/14	5/14
Overcast	4/9	**0/5**	Mild	4/9	3/5	Normal	6/9	1/5	F	6/9	2/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Likelihood of the two classes:

$$Pr(y = 1 | sunny, \ cool, \ high, \ T) \propto \frac{2}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{9}{14}$$

$$Pr(y = -1 | sunny, \ cool, \ high, \ T) \propto \frac{3}{5} \cdot \frac{1}{5} \cdot \frac{4}{5} \cdot \frac{3}{5} \cdot \frac{5}{14}$$

Zero-frequency Problem

- What if an attribute value does *NOT* occur with a class value?
  - The *posterior probability* will all be *zero*! No matter how likely the other attribute values are!
  - Laplace estimator will fix "zero-frequency", $\dfrac{k+\lambda}{n+a\lambda}$
- **Question:** Roll a dice 8 times. The outcomes are as: 2, 5, 6, 2, 1, 5, 3, 6. What is the probability for showing 4?

$$Pr(X=4)=\frac{0+\lambda}{8+6\lambda},\quad Pr(X=5)=\frac{2+\lambda}{8+6\lambda}$$