Machine Learning (I): Classification and Conditional Distribution

Le Wang

Motivation

We are interested in whether or not the relationship exists. But more important, we are interested in predictions.

Given a value of X, what will Y be?

Certainly the joint distribution is useful for informing whether or not the relationship between X and Y exists, but it does not tell us the answer to this question. We need something more straightforward to answer this question.

Necessary Definitions

More formally,

1. **Inputs** X: measured or present variables. Synonyms: predictors, features or independent variables - These inputs have some influence on one or more outputs.
2. **Output** Y is also called response or dependent variable or outcome variables.

Eventually we will try to learn the correspondence between X and Y:

$$Y = f(X)$$

Statistcal Learning: Supervised vs Unsupervised Learning

1. **Supervised Learning**: Presence of the outcome variable to guide the learning process (We have Y and X)

Goal: e.g. to use the inputs to predict the values of the outputs
Methods: regression methods (linear, lasso, ridge, etc.), bagging, trees, random forests, ensemble learning, . . .

Statistcal Learning: Supervised vs Unsupervised Learning

1. **Supervised Learning**: Presence of the outcome variable to guide the learning process (We have Y and X)

Goal: e.g. to use the inputs to predict the values of the outputs
Methods: regression methods (linear, lasso, ridge, etc.), bagging, trees, random forests, ensemble learning, ...

2. **Unsupervised Learning**: only features are observed, no measurements of the outcome variable (We have X, but not Y)

Goal: insights how the data are organized or clustered Methods: Association Rules, PCA, cluster analysis.

Stotistical Learning: What to Learn

General Goal: There are so many different values of Y. What to learn?

1. Distribution
2. When it is impossible to learn the entire distribution, we learn features or parts of the distribution.

Statistical Learning: Misconception

Regression vs Classification

1. Input variables X
2. Regression: **Quantitative** (continuous) output
3. Classification: **Qualitative** output (categorical / discrete)

Wrong type of ways to organize the methods! They are learning different things!

Statistical Learning: Classification Problems

We will discuss the case of **discrete** Y *and* **discrete** X. In this case, we can learn about the entire distribution of Y, which is completely **nonparametric** and model-free.

The case of discrete Y is closely related to the **classification problem** in machine learning. Chapter 4 in *An Introduction to Statistical Learning: with Applications in R*

Classification Problems

1. **Medical** A person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of three medical conditions. Which of the three conditions does the individual have?
2. **Finance** An online banking service must be able to determine whether or not a transaction being performed on the site is fraudulent, on the basis of the user's IP address, past transaction history, and so forth.
3. **Biology** On the basis of DNA sequence data for a number of patients with and without a given disease, a biologist would like to figure out which DNA mutations are deleterious (disease-causing) and which are not.

Classification Problems (-cont.-)

4. **Political Science** Whether or not a politician may win an election (given his/her characteristics and voter composition etc.)
5. **Sports** Whether or not a team will win a game given the characteristics of the team and its opponent, weather, and crowd, whether or not it is a home game.
6. **Computer Science** Your smart phone wants to predict your locations (home, office, restaurant, or store) based on the time of a day.

Classification Problems (-cont.-)

Approaches for predicting qualitative responses, a process that is known as **classification**. Predicting a qualitative response for an observation can be referred to as classifying that observation, since it involves assigning the observation to a category, or class.

Often the methods used for classification are called **classifiers**, typically involving the following steps:

Classification Problems (-cont.-)

Approaches for predicting qualitative responses, a process that is known as **classification**. Predicting a qualitative response for an observation can be referred to as classifying that observation, since it involves assigning the observation to a category, or class.

Often the methods used for classification are called **classifiers**, typically involving the following steps:

1. first predict the probability of each of the categories of a qualitative variable

Classification Problems (-cont.-)

Approaches for predicting qualitative responses, a process that is known as **classification**. Predicting a qualitative response for an observation can be referred to as classifying that observation, since it involves assigning the observation to a category, or class.

Often the methods used for classification are called **classifiers**, typically involving the following steps:

1. first predict the probability of each of the categories of a qualitative variable
2. based on the probabilities, make the classification.

Classification Problems: A Numerical Example

ID	X	Y
1	1	0
2	1	0
3	1	0
4	2	1
5	2	0
6	2	1
7	2	1

Questions: What are your predictions of Y when $X = 1, 2$, respectively?

Conditional Distributions

Definition. Conditional Distribution is a probability distribution for a sub-population. That is, a conditional probability distribution describes the probability that a randomly selected person from a sub-population has the one characteristic of interest.

$$\Pr[Y|X = x]$$

Conditional Distributions

Definition. Conditional Distribution is a probability distribution for a sub-population. That is, a conditional probability distribution describes the probability that a randomly selected person from a sub-population has the one characteristic of interest.

$$\Pr[Y|X = x]$$

Our Example:

1. $\Pr[Y|X = 1]$: $\Pr[Y = 0 \mid X = 1]$ and $\Pr[Y = 1 \mid X = 1]$
2. $\Pr[Y|X = 2]$: $\Pr[Y = 0 \mid X = 2]$ and $\Pr[Y = 1 \mid X = 2]$

Conditional Distribution (from Joint Distribution)

X/Y	1	2	3	4	5	6	$p(x_i)$
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
$p(y_i)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	

Conditional Distribution (from Joint Distribution)

What is the distribution of Y given $X = 1$

X/Y	1	2	3	4	5	6	$p(x_i)$
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$

Conditional Distribution (from Joint Distribution)

What is the distribution of Y given $X = 1$

X/Y	1	2	3	4	5	6	$p(x_i)$
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$

It would be $\frac{1}{36} \cdot N$ divided by $\frac{1}{6} \cdot N$.

Conditional Distribution (from Joint Distribution)

What is the distribution of Y given $X = 1$

X/Y	1	2	3	4	5	6	$p(x_i)$
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$

It would be $\frac{1}{36} \cdot N$ divided by $\frac{1}{6} \cdot N$.

It turns out that the information on the sample size is **NOT** required for calculation of the conditional distribution once we have the joint distribution.

Conditional Distribution (from Joint Distribution)

What is the distribution of Y given $X = 1$

X/Y	1	2	3	4	5	6	$p(x_i)$
1	$\frac{1}{36}/\frac{1}{6} = \frac{1}{6}$	$\frac{1}{36}/\frac{1}{6} = \frac{1}{6}$	$\frac{1}{36}/\frac{1}{6} = \frac{1}{6}$	$\frac{1}{36}/\frac{1}{6} = \frac{1}{6}$	$\frac{1}{36}/\frac{1}{6} = \frac{1}{6}$	$\frac{1}{36}/\frac{1}{6} = \frac{1}{6}$	

Conditional Distribution (from Joint Distribution)

What is the distribution of Y given $X = 1$

X/Y	1	2	3	4	5	6	$p(x_i)$
1	$\frac{1}{36}/\frac{1}{6} = \frac{1}{6}$	$\frac{1}{36}/\frac{1}{6} = \frac{1}{6}$	$\frac{1}{36}/\frac{1}{6} = \frac{1}{6}$	$\frac{1}{36}/\frac{1}{6} = \frac{1}{6}$	$\frac{1}{36}/\frac{1}{6} = \frac{1}{6}$	$\frac{1}{36}/\frac{1}{6} = \frac{1}{6}$	

Conditional Distribution

$$\Pr[Y \mid X] = \frac{\Pr[Y, X]}{\Pr[X]}$$

Conditional, Marginal, and Joint Distributions)

Conditional Distribution

$$\Pr[Y \mid X] = \frac{\Pr[Y, X]}{\Pr[X]}$$

is equivalent to

$$\Pr[Y, X] = \Pr[Y \mid X] \cdot \Pr[X]$$

Important We will use this equilvalent result to derive statistical language models.

Conditional Distribution: Super Bowl in R

Lets look at our example `mv03_cond_dist_superbowl.R`

Conditional Distribution, Prediction and Classification

Bayes classifier:

In this simple example with only two classes (values), the Bayes classifier generates the prediction

1. If $\Pr[Y = 0 \mid X = x_0] > 0.5$, then class $Y = 0$
2. If $\Pr[Y = 1 \mid X = x_0] > 0.5$, then class $Y = 1$

Conditional Distribution, Prediction and Classification

Bayes classifier:

In this simple example with only two classes (values), the Bayes classifier generates the prediction

1. If $\Pr[Y = 0 \mid X = x_0] > 0.5$, then class $Y = 0$
2. If $\Pr[Y = 1 \mid X = x_0] > 0.5$, then class $Y = 1$

Bayes classifier (general type): classify the **most probable** class

$$\max_y \Pr[Y = y \mid X = x_0]$$

Reasoning Behind Bayes Classifier

Error Rate: Percentage of errors that you make (where your forecast is $\widehat{y}$)

$$\mathbb{E}[\mathbb{I}[Y \neq \widehat{y}]]$$

How can I minimize the expected error rate?

Reasoning Behind Bayes Classifier

Suppose that the conditional distribution is as follows

$\Pr[Y = 0 \mid X = x_0] = .7$ and $\Pr[Y = 0 \mid X = x_0] = .3$

What is the error rate?

Reasoning Behind Bayes Classifier

Suppose that the conditional distribution is as follows

$\Pr[Y = 0 \mid X = x_0] = .7$ and $\Pr[Y = 0 \mid X = x_0] = .3$

What is the error rate?

If $\widehat{y} = 0$, $\mathbb{E}[\mathbb{I}[Y \neq 0]] = \Pr[Y = 1 \mid X = x_0] = .3$

Reasoning Behind Bayes Classifier

Suppose that the conditional distribution is as follows

$\Pr[Y = 0 \mid X = x_0] = .7$ and $\Pr[Y = 0 \mid X = x_0] = .3$

What is the error rate?

If $\widehat{y} = 0$, $\mathbb{E}[\mathbb{I}[Y \neq 0]] = \Pr[Y = 1 \mid X = x_0] = .3$

If $\widehat{y} = 1$, $\mathbb{E}[\mathbb{I}[Y \neq 1]] = \Pr[Y = 1 \mid X = x_0] = .7$

Reasoning Behind Bayes Classifier

Suppose that the conditional distribution is as follows

$\Pr[Y = 0 \mid X = x_0] = .7$ and $\Pr[Y = 0 \mid X = x_0] = .3$

What is the error rate?

If $\widehat{y} = 0$, $\mathbb{E}[\mathbb{I}[Y \neq 0]] = \Pr[Y = 1 \mid X = x_0] = .3$

If $\widehat{y} = 1$, $\mathbb{E}[\mathbb{I}[Y \neq 1]] = \Pr[Y = 1 \mid X = x_0] = .7$

In summary,

$$\mathbb{E}[\mathbb{I}[Y \neq \widehat{y}]] = 1 - \Pr[Y = \widehat{y}]$$

Reasoning Behind Bayes Classifier

Expected Error Rate,

$$\mathbb{E}[\mathbb{I}[Y \neq \widehat{y}]] = 1 - \Pr[Y = \widehat{y}]$$

How to minimize this one?

Reasoning Behind Bayes Classifier

Expected Error Rate,

$$\mathbb{E}[\mathbb{I}[Y \neq \widehat{y}]] = 1 - \Pr[Y = \widehat{y}]$$

How to minimize this one?

Choose the one with the maximum $\Pr[Y = \widehat{y}]$

Bayes Classifier: Implementation in R

We will look at the single variable case where naive Bayes classifier coincides with the Bayes classifer to ease the implementation in R.

`mv03_cond_dist_naive-bayes.R`

Note Naive Bayes Classifer actually adds more assumptions when computing the conditional probabilities when we have multiple variables. We will introduce it later when we introduce the Bayes rule.

Extension to More than Two Variables

1. An approach based on the definition
2. An approach based on the link to the mean

Extension to More than Two Variables: Approach 1 based on Definition

$$\Pr[X \text{ and } Y|Z] = \frac{\Pr[X \text{ and } Y \text{ and } Z]}{\Pr[Z]}$$

$$\Pr[Y|X, Z] = \frac{\Pr[X \text{ and } Y \text{ and } Z]}{\Pr[X \text{ and } Z]}$$

$$\Pr[Y, X|Z, W] = \frac{\Pr[X \text{ and } Y \text{ and } Z \text{ and } W]}{\Pr[Z \text{ and } W]}$$

Note that it does not change our Bayes classifier. We can simply think of X, Z as a giant X.

Extension to More than Two Variables: Approach 1 based on Definition

$$\Pr[\text{Outcome} \mid \text{Predictor}] = \frac{\Pr[\text{Outcome}, \text{Predictor}]}{\Pr[\text{Predictor}]}$$

Intuitive Way: No many how many variables you have as outcome or predictor variables. Just think of them as one variable with $m_1 \times m_2 \times m_3 \cdots \times m_k$ values.

Extension to More than Two Variables: Approach 1 based on Definition

R code to implement the multiple-variable case.

Extension to More than Two Variables: Approach 2 based on Mean

Remember that

$$\Pr[Y = y] = \mathbb{E}[\mathbb{I}(Y = y)]$$

Extension to More than Two Variables: Approach 2 based on Mean

Remember that

$$\Pr[Y = y] = \mathbb{E}[\mathbb{I}(Y = y)]$$

$$\Pr[Y = y | X = x] = \mathbb{E}[\mathbb{I}(Y = y) | X = x]$$

For every value, y, of Y, we can generate an indicator variable, equal to one if $Y = y$, zero otherwise. These indicator variables are also called **dummy variables**.

For any categorical variables (factor variables in R), we can create a dummy variable for each category.

Step 1. If Y can take only four different values (say, $1, 2, 3, 4$ or first, second, third, fourth seasons), then we need to create four **additional** dummy variables, denoted by I_1, I_2, I_3, I_4:

$$I_1 = \mathbb{I}[Y = 1]$$
$$I_2 = \mathbb{I}[Y = 2]$$
$$I_3 = \mathbb{I}[Y = 3]$$
$$I_4 = \mathbb{I}[Y = 3]$$

*Note:** This is also a useful trick to consider more flexbile functions in estimations. We will learn how to generate such variables with the `factor()` and `model.matrix()` commands.

Step 2. We then calculate the **mean** of each of the four **new** variables for each category of $X = (x_1, x_2, x_3, \dots)$, which can be taken care of easily using the `aggregrate()` command.

Lets look at our example `/mv03_cond_dist_multiplevars02.R`

An Application to Consider: Airbnb

A match is determined by two sides: both guests and hosts

An Application to Consider: Airbnb

A match is determined by two sides: both guests and hosts

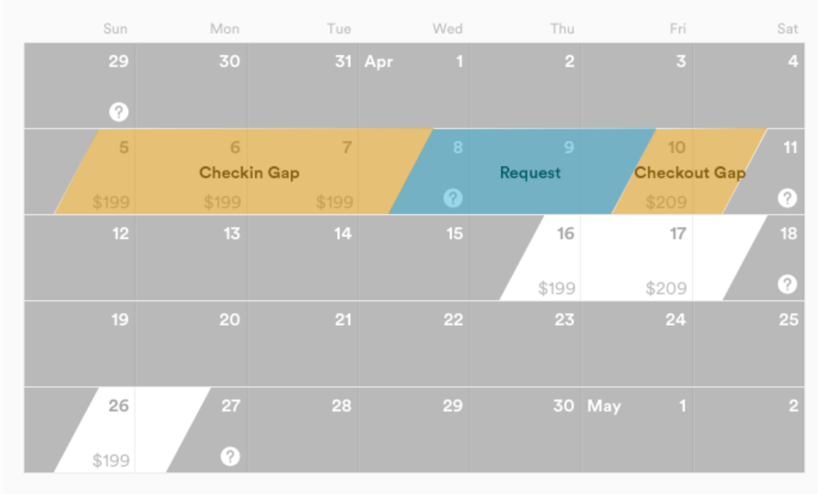
Understand Host Preferences, classify them into **acceptance** vs. **non-acceptance**

An Application to Consider: Airbnb

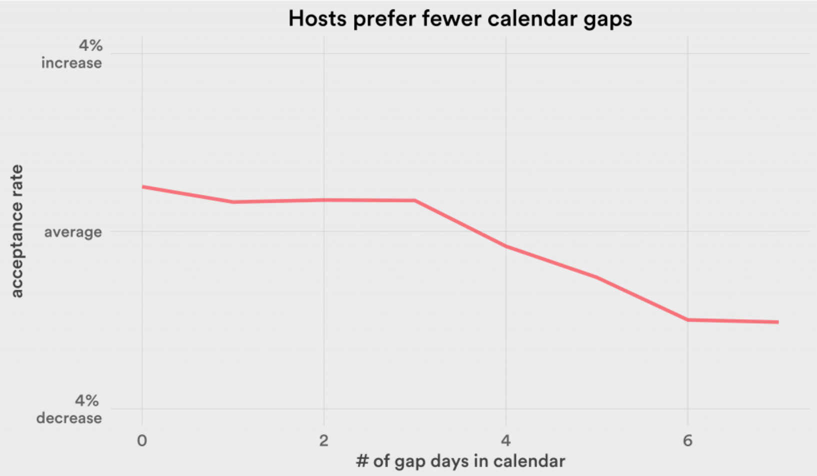
A match is determined by two sides: both guests and hosts

Understand Host Preferences, classify them into **acceptance** vs. **non-acceptance**

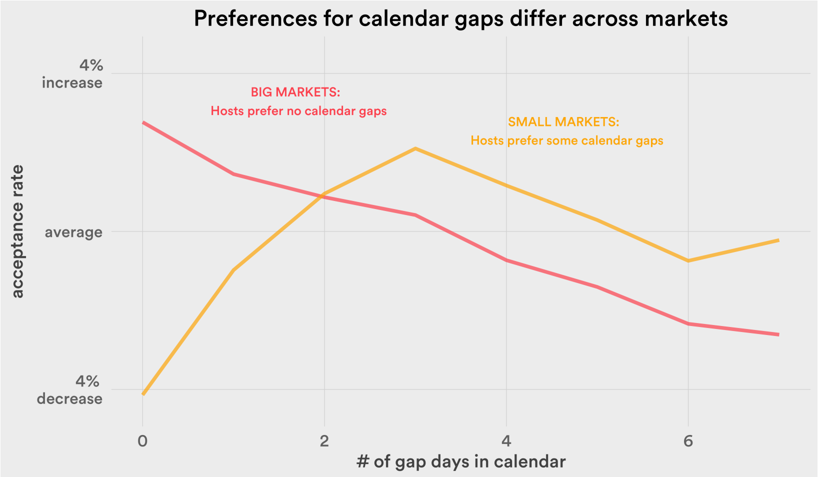
Question: Do they maximize the occupancy?



Sun
Mon
Tue
Wed
Thu
Fri
Sat
29
30
31
Apr
1
2
3
4
5
6
7
8
9
10
11
Checkin Gap
Request
Checkout Gap
$199
$199
$199
$209
12
13
14
15
16
17
18
$199
$209
19
20
21
22
23
24
25
26
27
28
29
30
May
1
2
$199


Hosts prefer fewer calendar gaps
4%
increase
average
4%
decrease
acceptance rate
0
2
4
6
# of gap days in calendar

further market size: Heterogeneity in Host Preferences



For more detailed discussions available here