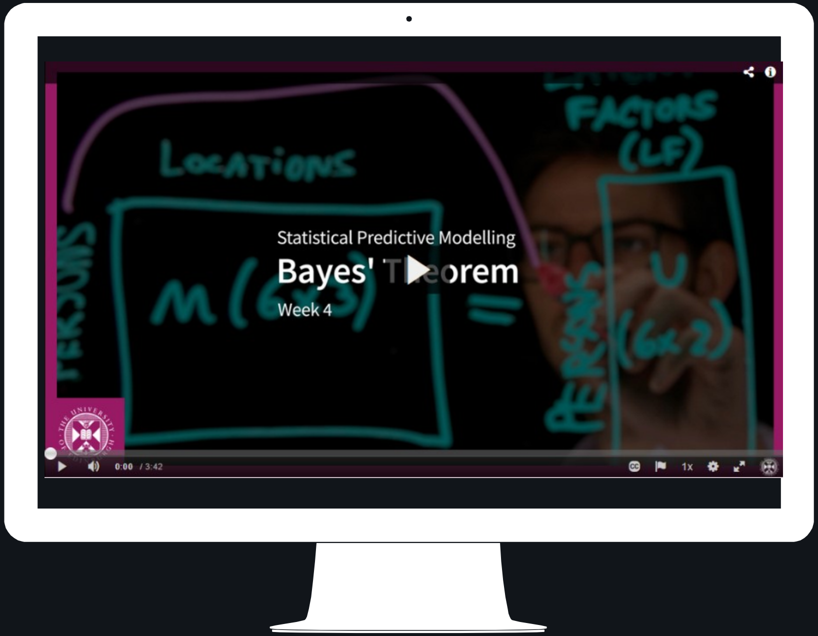
UNIVERSITY OF EDINBURGH
Business School

Predictive Analytics and Modelling of Data

CMSE11428 (2020-2021)

Dr Xuefei Lu
The University of Edinburgh Business School

Introduction to Naïve Bayes



Bayes' Theorem

- Please watch the following video via this link
- https://media.ed.ac.uk/media/Bayes%27+Theorem/1_25qb6kfw/122596071

Basic concepts of probability

Probability is used to describe or measure the uncertainty associated with the outcome of a future experiment.

- The 'experiment': the number on the roll of a dice.
- Each of these outcomes or events may have a probability associated with it, which is the number from 0 (will not occur) to 1 (certain to occur).

The frequentist approach: requires counting the number of times the event has occurred (r) in a large number of trials (n). The probability, denoted by 'P' is then the relative frequency of the event:

$$P = \frac{r}{n},$$

The odds that event A will occur are given by the ratio of the probability that it will occur to the probability that it will not occur:

$$\mathrm{Odds(A)} = \frac{P(A)}{1 - P(A)}, \qquad \text{and } \mathrm{P(A)} = \frac{\mathrm{Odds(A)}}{\mathrm{Odds(A)} + 1}$$

If a event only have two outcomes -- event occurs or not, their probabilities sum up to 1.

https://dribbble.com/shots/2479300-Thirteen-gif

Basic concepts of probability: Conditional Probability

We are interested in how the probability of one event (A) is affected by another event (B).

The probability of A given B:

$$P\ (A \mid B) = \frac{P\ (A \cap B)}{P(B)}$$

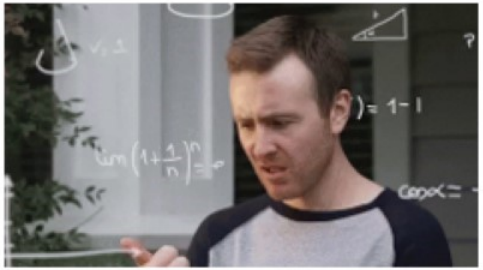
If A and B are independent of each another, then $P\ (A \mid B) = P(\mathrm{A})$ -- the chances of A happening are not affected by the fact that B has happened or not. E.g. the outcome of rolling a dice.

The probability of A happening given B_1, B_2 and B_3 has happened: $P\ (A|B_1, B_2, B_3)$.

Basic concepts of probability: retail credit example

Suppose we have an online Fintech company that starts lending money. Let's call it Credit Miracle. In the first year of its operation it lends to 1,000 customers, and does not reject anyone. At the end of the year, it turns out that 700 out of 1000 have turned out to be Good customers, that is, paid back their loans. The remaining 300 have not paid back, so we label them as Bad.

Q1: What is the estimated probability of being a Bad customer for anyone who will apply next year? You can assume that the future is going to be like the past, no changes.



https://www.datashieldprotect.com/blog/is-cyber-insurance-reliable

Basic concepts of probability: retail credit example

Suppose we have an online Fintech company that starts lending money. Let’s call it Credit Miracle. In the first year of its operation it lends to 1,000 customers, and does not reject anyone. At the end of the year, it turns out that 700 out of 1000 have turned out to be Good customers, that is, paid back their loans. The remaining 300 have not paid back, so we label them as Bad.

Q1: What is the estimated probability of being a Bad customer for anyone who will apply next year? You can assume that the future is going to be like the past, no changes.

Correct answer: 0.3 or 30%.
Based on historic numbers, there are 300 Bads out of 1000 total, so 300/1000=0.3.

Basic concepts of probability: retail credit example

Suppose we have an online Fintech company that starts lending money. Let's call it Credit Miracle. In the first year of its operation it lends to 1,000 customers, and does not reject anyone. At the end of the year, it turns out that 700 out of 1000 have turned out to be Good customers, that is, paid back their loans. The remaining 300 have not paid back, so we label them as Bad.

Q2: What would be the probability of a new credit applicant becoming a Good customer?

Basic concepts of probability: retail credit example

Suppose we have an online Fintech company that starts lending money. Let's call it Credit Miracle. In the first year of its operation it lends to 1,000 customers, and does not reject anyone. At the end of the year, it turns out that 700 out of 1000 have turned out to be Good customers, that is, paid back their loans. The remaining 300 have not paid back, so we label them as Bad.

Q2: What would be the probability of a new credit applicant becoming a Good customer?

Correct answer: 0.7 or 70%. Based on historic numbers, there are 700 Goods out of 1000 total, so 700/1000=0.7. Or knowing that probability of being Bad, we can simply subtract it from 1, because the probabilities of all possible outcomes must sum to 1.

https://giphy.com/gifs/d8isjk1UBPFTm0EBbd/html5

Basic concepts of probability: retail credit example

Credit Miracle is not happy with this Probability of Default (PD), and would like to reduce it. If only they could identify the Bads when they apply for loans and reject them. Unfortunately, this cannot be observed directly. Additional information: if the credit applicant gave their telephone number when applying for credit.

The **contingency, or cross-tabulation**:

Telephone	Status		All
	Bad	Good	
Known (Yes)	113	291	404
Unknown (No)	187	409	596
All	300	700	1000

The probability of not knowing the phone number for Bad customers (given Bad):

$$P(No|\,Bad) = \frac{P(No \cap Bad)}{P(Bad)} = \frac{\frac{187}{1000}}{\frac{300}{1000}} = 0.62$$

Basic concepts of probability: retail credit example

Credit Miracle is not happy with this Probability of Default (PD), and would like to reduce it. If only they could identify the Bads when they apply for loans and reject them. Unfortunately, this cannot be observed directly. Suppose there is other information we can use.

Telephone	Status		All
	Bad	Good	
Known (Yes)	113	291	404
Unknown (No)	187	409	596
All	300	700	1000

Q3: Calculate the probability for not giving the phone number 'given Good'. You can use a calculator or a piece of paper.

Basic concepts of probability: retail credit example

Telephone	Status		All
	Bad	Good	
Known (Yes)	113	291	404
Unknown (No)	187	409	596
All	300	700	1000

Q3: Calculate the probability for not giving the phone number 'given Good'.

Correct answer: 0.58 or 58%.

The total now changes from 'all customers' to 'Goods' because we are interested in the conditional probability. Based on historic numbers, there are 409 Goods with unknown telephone out of 700 total Goods, so 409/700=0.58.

P(known telephone | Good) + P(unknown telephone| Good) = 1

Basic concepts of probability: retail credit example

Telephone	Status		All
	Bad	Good	
Known (Yes)	113	291	404
Unknown (No)	187	409	596
All	300	700	1000

The probability of not knowing the phone number for Bad customers (given Bad):62%

The probability for not giving the phone number 'given Good': 58%.

Q4: What is your conclusion? Is there a difference between Good and Bad customers?

Basic concepts of probability: retail credit example

Telephone		Status	All
	Bad	Good	
Known (Yes)	113	291	404
Unknown (No)	187	409	596
All	300	700	1000

The probability of not knowing the phone number for Bad customers (given Bad): 62%

The probability for not giving the phone number 'given Good': 58%.

Q4: What is your conclusion? Is there a difference between Good and Bad customers?

https://www.pinterest.co.uk/pin/630152172845000888/

Basic concepts of probability: retail credit example

Q4: What is your conclusion? Is there a difference between Good and Bad customers?

YES: The conditional probabilities are given below in the Contingency Table

Telephone	Status (Conditional on)		All
	Bad	Good	
Known (Yes)	0.38	0.42	0.40
Unknown (No)	0.62	0.58	0.60

Understanding Bayes

In the context of our credit example:

- $P\ (G)$ is the probability that a borrower will be a Good customer (i.e. will repay the money borrowed), or the proportion of Goods in the portfolio.
- $P\ (B)$ is the probability that a borrower will be a Bad customer, or the proportion of Bads in the portfolio or portfolio segment

$$P\ (G)\ +\ P\ (B)\ = 1$$

- Let $(X_1, X_2, \ldots, X_p)$ be characteristics (variables) of the borrower such as age, marital status, housing, etc. And $\boldsymbol{x} = \ (x_1, x_2, \ldots, x_p)$ be outcomes/realisation of characteristics of the borrower.
- $P\ (\boldsymbol{x})$ is the probability of observing a certain attribute -- marginal, since it appears on the margin of the contingency table.
- The probability $P\ (\cdot)$ is often referred to as **likelihood**, e.g. $P\ (B)$ and $P\ (G)$ they indicate how likely a certain attribute is in Good or Bad class.

Understanding Bayes

In Bayesian notation $P(G)$ and $P(B)$ are **prior probabilities**, since these are the probabilities that we started with and we assume to know them in priori.

Then we want to use new information on $\boldsymbol{x}$ to obtain our main interest or objective:

- $P(G|\boldsymbol{x})$ is the probability that a borrower will be a Good customer (i.e. will repay the money borrowed on conditions specified) given certain attributes;
- $P(B|\boldsymbol{x})$ is the probability that a borrower will be a Bad customer given certain attributes.

We are interested in estimating **posterior probability** of default (being Bad) given certain characteristics, we can use the **Bayes Rule**:

$$P(B|\boldsymbol{x}) = \frac{P(B)P(\boldsymbol{x}|B)}{P(\boldsymbol{x})}$$

Understanding Bayes

The **Bayes Rule**:

$$P(B|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|B)P(B)}{P(\boldsymbol{x})}$$

Proof:

By definition of conditional probability, we have

$$P(B|\boldsymbol{x}) = \frac{P(B\cap\boldsymbol{x})}{P(\boldsymbol{x})} \text{ and } P(\boldsymbol{x}|B) = \frac{P(\boldsymbol{x}\cap B)}{P(B)}$$

From the second definition, we have $P(\boldsymbol{x} \cap B) = P(\boldsymbol{x}|B)P(B)$.

Note. that $P(B \cap \boldsymbol{x}) = P(\boldsymbol{x} \cap B)$

Therefore $P(B|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|B)P(B)}{P(\boldsymbol{x})}$.

Understanding Bayes

Back to our Credit Miracle example. we calculated priors: $P\ (B)\ =\ 0.3,\ P(G) = \ 0.7$, and for the variable 'Telephone', marginal probabilities and likelihoods:

Telephone $\boldsymbol{x}$	Status $P(\boldsymbol{x}\|B)$		$P(\boldsymbol{x})$
	Bad	Good	
Known (Yes)	0.38	0.42	0.40
Unknown (No)	0.62	0.58	0.60

$P(known) = \ 0.40, P(unknown) = \ 0.60$

$P(known|Bad) = \ 0.38, P(known|Good) = \ 0.416$

$P(unknown|Bad) = \ 0.623, P(known|Good) = \ 0.584$

We are really interested in posterior probabilities for any future applicants of being Bad or Good if they provide a telephone number: $P(Bad|unknown)$ and $P(Good|unknown)$

Understanding Bayes

$P\ (B)\ =\ 0.3, P(G) =\ 0.7, P(known) =\ 0.40, P(unknown) =\ 0.60$

$P(known|Bad) =\ 0.38, P(known|Good) =\ 0.42$

$P(unknown|Bad) =\ 0.62, P(unknown|Good) =\ 0.60$

We are really interested in posterior probabilities for any future applicants of being Bad or Good if they provide a telephone number: $P(Bad|unknown)$ and $P(Good|unknown)$.

If telephone known:

$$P(Bad|known) = \frac{P(known|B)P(B)}{P(known)} = \frac{0.38 \times 0.3}{0.40} = 0.28,$$

$$P(Good|known) = \frac{P(known|good)P(G)}{P(known)} = \frac{0.42 \times 0.7}{0.40} = 0.72,$$

If telephone unknown:

$$P(Bad|unknown) = \frac{P(unknown|B)P(B)}{P(unknown)} = \frac{0.62 \times 0.3}{0.60} = 0.31,$$

$$P(Good|unknown) = \frac{P(unknown|good)P(G)}{P(unknown)} = \frac{0.60 \times 0.7}{0.60} = 0.70,$$

Understanding Bayes

If telephone known:

$$P(Bad|known) = \frac{P(known|B)P(B)}{P(known)} = 0.28,$$

$$P(Good|known) = \frac{P(known|good)P(G)}{P(known)} = 0.72,$$

If telephone unknown:

$$P(Bad|unknown) = \frac{P(unknown|B)P(B)}{P(unknown)} = 0.31,$$

$$P(Good|unknown) = \frac{P(unknown|good)P(G)}{P(unknown)} = 0.70,$$

The way that **Naive Bayes** works is by calculating the posterior probability of belonging to a certain class, and then assigning the observation to a class with the highest posterior probability.

We should assign both known and unknown categories to the Good class:

- 0.720 > 0.280, therefore we assign an applicant who gives a telephone number to a Good class;
- 0.686> 0.314, therefore we assign an applicant who does not give a telephone number to a Good class too.

Understanding Bayes

If telephone known:

$$P(Bad|known) = \frac{P(known|B)P(B)}{P(known)} = 0.28,$$

$$P(Good|known) = \frac{P(known|good)P(G)}{P(known)} = 0.72,$$

If telephone unknown:

$$P(Bad|unknown) = \frac{P(unknown|B)P(B)}{P(unknown)} = 0.31,$$

$$P(Good|unknown) = \frac{P(unknown|good)P(G)}{P(unknown)} = 0.70,$$

We can also change the decision rule:

If posterior probability is above $P\ (B)\ =\ 0.3$ – assign the category to a Bad class;

If below – assign the category to a Good class.

Using this rule 'known' should be classified as Good, and 'unknown' – as Bad.

Telephone	Status		All
	Bad	Good	
Known (Yes)	113	291	404
Unknown (No)	187	409	596
All	300	700	1000

Please study the following files:

- 10 - Activity 5 - Calculate conditional probabilities.ipynb + german_credit.csv

Then try the following exercise:

- 11 - Activity 6 - Explore predictive accuracy.ipynb + german.csv

Activity: Calculate conditional probabilities and predictions

Improving the accuracy of your predictions

A 2-way contingency table for 'Telephone			
Telephone Known?	Status		All
	Good	Bad	
No	409	187	596
Yes	291	113	404
All	700	300	1000

A 2-way contingency table for 'Housing'			
Housing	Status		All
	Good	Bad	
Rent	109	70	179
Own	527	186	713
Free	64	44	108
All	700	300	1000

	Status							
	Good				Bad			
Housing	Rent	Own	Free	All	Rent	Own	Free	All
Telephone								
No	66	319	24	409	50	114	23	187
Yes	43	208	40	291	20	72	21	113
All	109	527	64	700	70	186	44	300

Improving the accuracy of your predictions

	Status							
	Good				Bad			
Housing	Rent	Own	Free	All	Rent	Own	Free	All
Telephone								
No	66	319	24	409	50	114	23	187
Yes	43	208	40	291	20	72	21	113
All	109	527	64	700	70	186	44	300

$$P(Bad|\ No\ Phone\ \cap\ \ Rent)\ =\ 50/(66+50)\ \ =\ 0.43$$
$$P(No\ Phone\ \cap\ \ Rent|\ Bad) =\ 50/300\ =\ 0.167$$

Check if Bayes theorem holds

$$Posterior P(class|data) = \frac{Prior P(class) \times Likelihood P(data|class)}{P(data)}$$

$$P(Bad|NoPhone \cap Rent) = \frac{P(B) \times P(NoPhone \cap Rent|Bad)}{P(NoPhone \cap Rent)} = \frac{0.3 \times 0.167}{(66+50)/1000} = 0.43$$

Improving the accuracy of your predictions

	Status							
	Good				Bad			
Housing	Rent	Own	Free	All	Rent	Own	Free	All
Telephone								
No	66	319	24	409	50	114	23	187
Yes	43	208	40	291	20	72	21	113
All	109	527	64	700	70	186	44	300

However, if we continue adding predictors, this approach becomes very messy, since each new variable will bring an additional dimension to our contingency table.

The solution is to assume that predictors are independent given Status, this is called **conditional independence.** The joint probability of independent events is the product of individual probabilities.

So, for Housing and Telephone, we can multiply two individual likelihoods coming from individual frequency tables.

$$P(NoPhone \cap Rent|Bad) = P(NoPhone|Bad) \times P(Rent|Bad) = \frac{187}{300} \times \frac{70}{300} = 0.623 \times 0.233 = 0.145$$

This is close to the actual result of 0.167.

This assumption simplifies the estimation considerably: no need to look at all possible combinations.

It may be 'naive', since, in reality, predictors are rarely independent.

Improving the accuracy of your predictions

	Status							
	Good				Bad			
Housing	Rent	Own	Free	All	Rent	Own	Free	All
Telephone								
No	66	319	24	409	50	114	23	187
Yes	43	208	40	291	20	72	21	113
All	109	527	64	700	70	186	44	300

For the denominator of Bayes' formula, we need $P(data)$ or in our case $P(NoPhone \cap Rent)$. Since there are only two possibilities (each customer can only be Bad or Good), we can calculate it as:

$$P(NoPhone \cap Rent) = P(Bad) \times P(NoPhone \cap Rent|Bad) + P(Good) \times P(NoPhone \cap Rent|Good)$$

$$P(NoPhone \cap Rent|Good) = P(NoPhone|Good) \times P(Rent|Good) = \frac{409}{700} \times \frac{109}{700} = 0.584 \times 0.156 = 0.091$$

So we have calculated all components needed for estimation of $P(Bad|NoPhone \cap Rent)$ from individual frequency tables without referring to the 3-way one

$$P(Bad|NoPhone \cap Rent) = \frac{0.3 \times 0.145}{0.3 \times 0.145 + 0.7 \times 0.091} = 0.401$$

Please note that we CANNOT multiply posterior probabilities from each predictor to get a joint one.
➔ Bayes theorem.

$$P(Bad|NoPhone \cap Rent) \neq P(Bad|NoPhone) \times P(Bad|Rent)$$

UNIVERSITY OF EDINBURGH
Business School