Féidearthachtaí as Cuimse
Infinite Possibilities

Machine Learning

Lecture 4: Probability-based Learning (Naïve Bayes)

Bojan Božić & Bujar Raufi
School of Computer Science
TU Dublin, Grangegorman

bojan.bozic@tudublin.ie; bujar.raufi@tudublin.ie

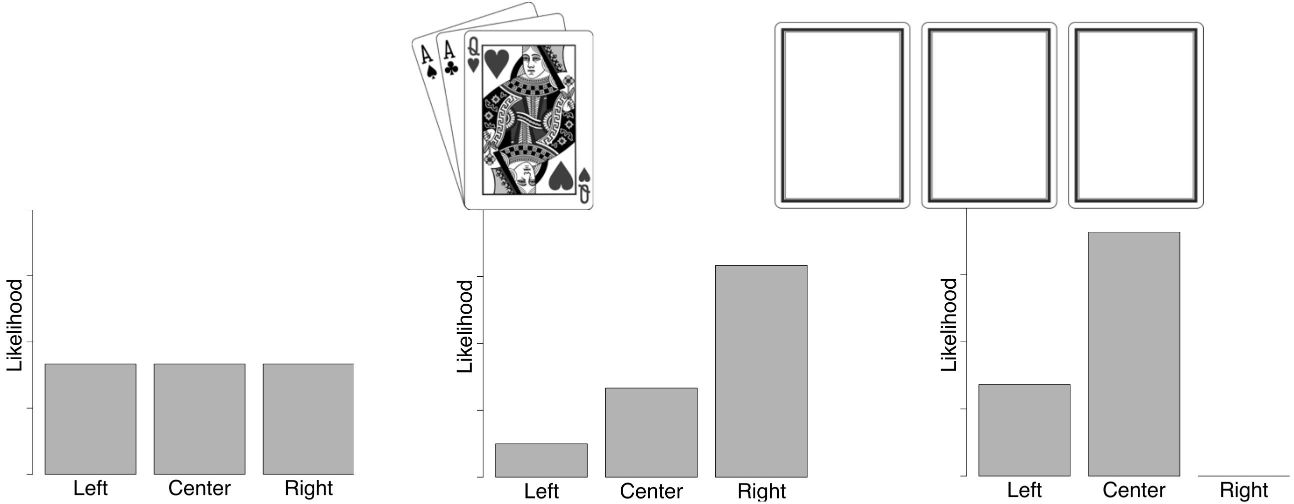
OLLSCOIL TEICNEOLAÍOCHTA BHAILE ÁTHA CLIATH
DUBLIN
TECHNOLOGICAL UNIVERSITY DUBLIN

Slides adapted from Sarah Jane Delany and book slides from: Fundamentals of Machine Learning for Predictive Data Analytics. Kelleher, Mac Namee and D'Arcy

Overview

- Big Idea
- Probability-based Learning
- Fundamentals
- Bayes Theorem
- Bayes Classification
- Smoothing, Binning, Continuous Features

Big Idea



(a) Initial likelihoods (b) Likelihoods after evidence (c) Revised likelihoods

- Probability based learning uses estimates of likelihoods to determine the most likely predictions that should be made
- Estimates are revised based on the data collected

Probability-based Learning

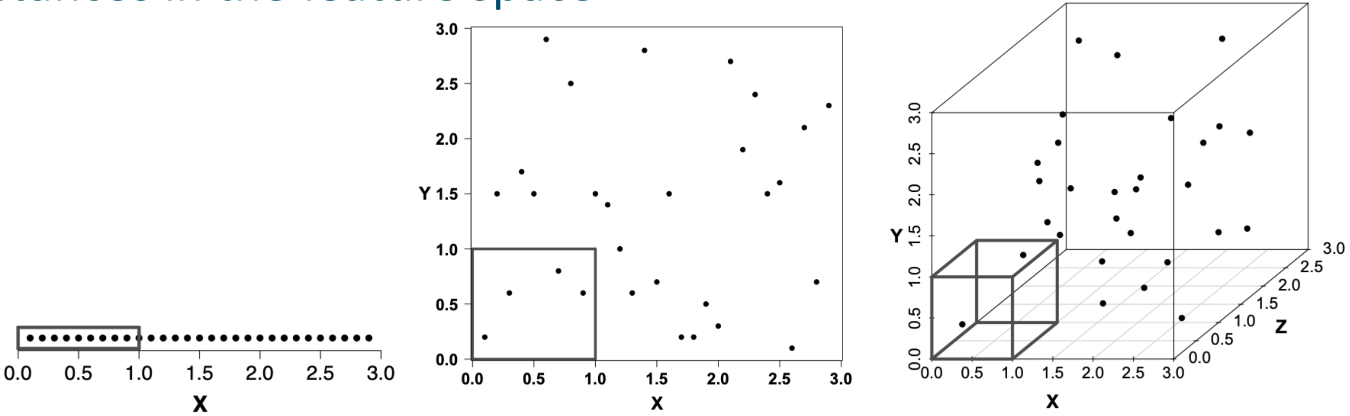
- Most common probabilistic approach to prediction is the Naïve Bayes classifier, an eager based learning approach based on Bayes Theorem

- Advantages of Naive Bayes
  - Simple and quick to train
  - Can handle large datasets
  - Good on sparse datasets (text)
  - Can handle missing values

- Although not as powerful as some other prediction models, they provide reasonable accuracy results while being robust to the curse of dimensionality and easy to train

Aside: The Curse of Dimensionality

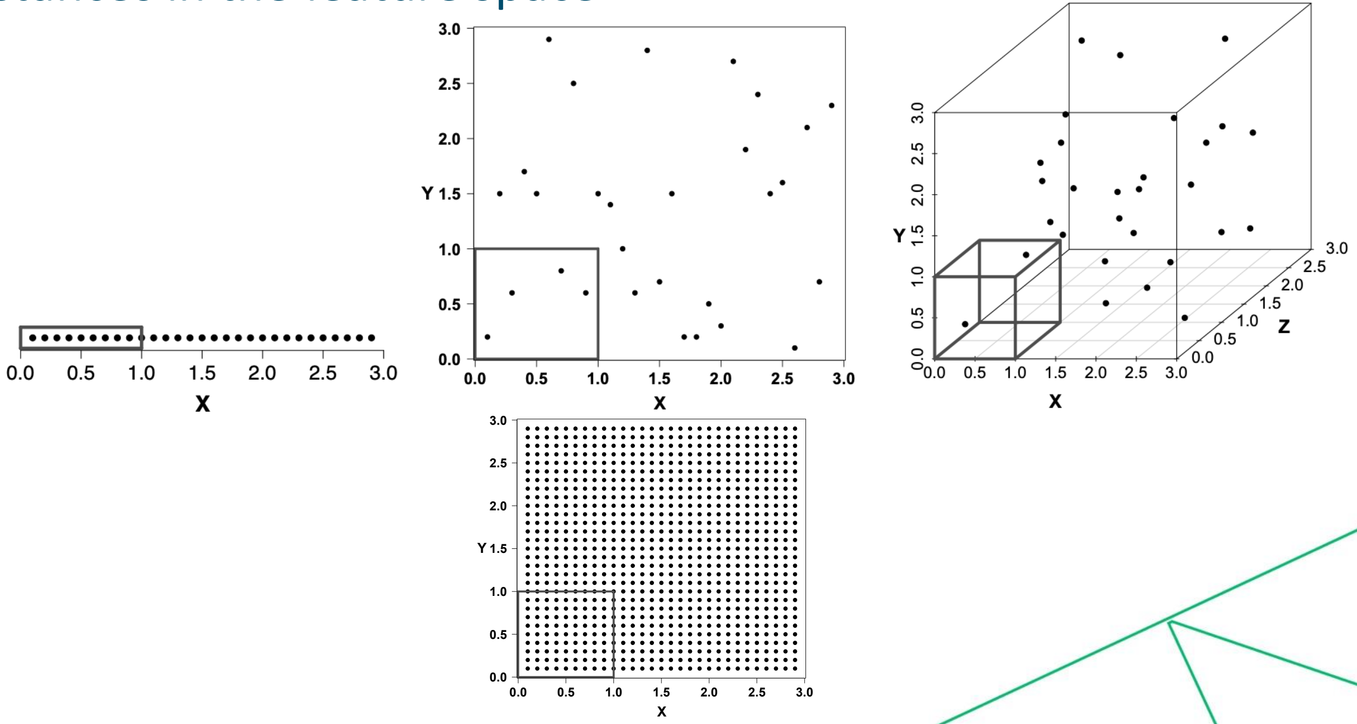
- Trade off between the number of descriptive features and the density of instances in the feature space



- To maintain the sampling density of the feature space as the number of descriptive features increase we need to dramatically increase the number of instances

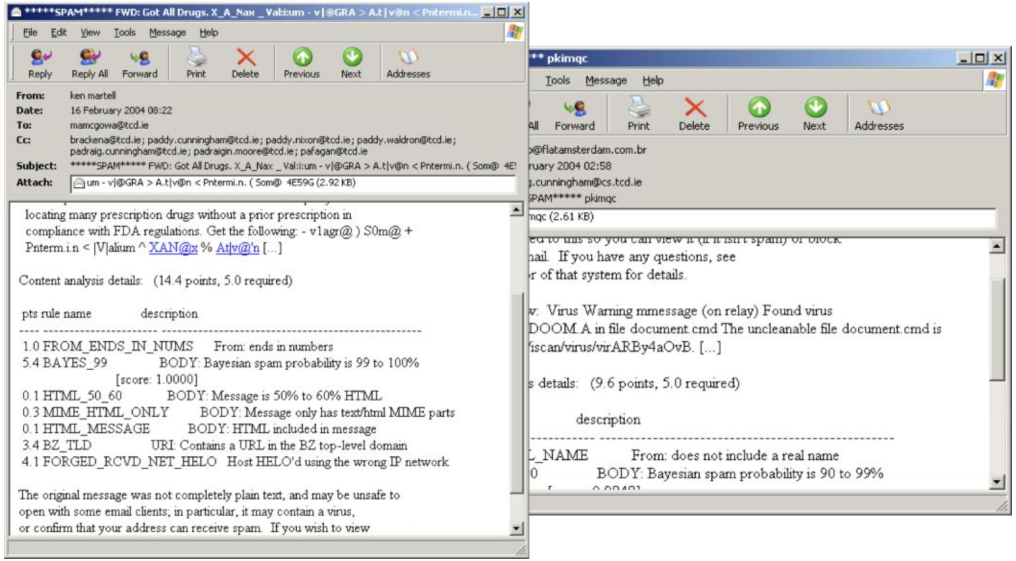
Aside: The Curse of Dimensionality

- Trade off between the number of descriptive features and the density of instances in the feature space



Application: SpamAssassin

- Apache SpamAssassin uses Naïve Bayes classification.



- See: http://wiki.apache.org/spamassassin/BayesInSpamAssassin

Fundamentals

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

- A probability function *P( )* returns the probability of an event (e.g. a feature taking a particular value)
- A joint probability - the prob of an assignment of specific values to multiple different features
- A conditional probability - the prob of one feature taking a specific value given we know the value of a different feature

Fundamentals

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

$$P(\text{Fever}) = ?$$

$$P(\text{Menigitis}, \text{Headache}) = ?$$

$$P(\text{Menigitis} | \text{Headache}) = ?$$

- A probability function *P( )* returns the probability of an event (e.g. a feature taking a particular value)
- A joint probability - the prob of an assignment of specific values to multiple different features
- A conditional probability - the prob of one feature taking a specific value given we know the value of a different feature

OLLSCOIL TEICNEOLAÍOCHTA BHAILE ÁTHA CLIATH
TU DUBLIN
TECHNOLOGICAL UNIVERSITY DUBLIN

Thomas Bayes

- 1701 - 1761
- Presbyterian minister
- Bayes Theorem published after his death

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

Probability that an event has happened given a set of evidence = the probability of the evidence being caused by the event x the probability of the event itself

Reasoning from evidence to event (inverse reasoning) is more difficult than reasoning from event to the evidence it causes (forward reasoning)

Fun fact: this is almost certainly not a picture of him.

https://www.york.ac.uk/depts/maths/histstat/bayespic.htm

OLLSCOIL TEICNEOLAÍOCHTA BHAILE ÁTHA CLIATH
TU DUBLIN
TECHNOLOGICAL UNIVERSITY DUBLIN

Example

- 20 lectures in a module, you attend 15
- 4 wet days, you attend on 2 of these

	Attend	Miss	
Wet			
Dry			

- P(A|W) = ?
- P(W|A) = ?

Example using Bayes Theorem

Example

After a yearly checkup, a doctor informs their patient that he has both bad news and good news. The bad news is that the patient has tested positive for a serious disease and that the test that the doctor has used is 99% accurate (i.e., the probability of testing positive when a patient has the disease is 0.99, as is the probability of testing negative when a patient does not have the disease). The good news, however, is that the disease is extremely rare, striking only 1 in 10,000 people.

- What is the actual probability that the patient has the disease?
- Why is the rarity of the disease good news for the patient given that she has tested positive?

OLLSCOIL TEICNEOLAÍOCHTA BHAILE ÁTHA CLIATH
TU DUBLIN
TECHNOLOGICAL UNIVERSITY DUBLIN

Example

$$P(d|t) = \frac{P(t|d)P(d)}{\color{red}{P(t)}}$$

$$\begin{aligned} P(t) &= P(t|d)P(d) + P(t|\neg d)P(\neg d) \\ &= (0.99 \times 0.0001) + (0.01 \times 0.9999) = 0.0101 \end{aligned}$$

$$\begin{aligned} P(d|t) &= \frac{0.99 \times 0.0001}{0.0101} \\ &= 0.0098 \end{aligned}$$

Why is the rarity of the disease good news?

Bayes Theorem

"The probability that an event has happened given a set of evidence for it is equal to the probability of the evidence being caused by the event by the probability of the event itself."

What is the probability of a given hypothesis *h* being true ("the event"), given the observed data *D* ("the evidence")?

Bayes Theorem: Rule states that for each possible hypothesis *h*

Likelihood

Prior Probability of the Hypothesis

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Posterior Probability

Prior Probability of the Data

Example: Bayes Theorem

D: Helen is 28 years old, is on a bill-pay plan, and earns €40k.

h: Helen will buy a new iPhone

$P(h\|D)$ Posterior Probability of *h*	Probability that Helen will buy a new iPhone, given that we know her age, plan, and income.
$P(h)$ Prior Probability of *h*	Probability that Helen will buy a new iPhone regardless of age, plan, and income
$P(D\|h)$ Posterior Probability of *D*	Probability that Helen is 28 years old, is on a bill-pay plan, and earns €40k, given that she has bought the iPhone 8.
$P(D)$ Prior Probability of *D*	Probability that a person from our dataset of customers is 28 years old, is on a bill-pay plan, and earns €40k.

We can calculate the Posterior Probability of *h* using Bayes Theorem:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Example: Bayes Theorem

- In the training set for a spam email filtering system:
  - 30 out of 74 emails are marked as spam
  - 51 emails of those 74 contain the word "free"
  - 20 emails containing the word "free" are marked as spam
- ***h***: Is a new email spam, given that it contains the word "free"?

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \qquad P(\text{spam}|\text{free}) = \frac{P(\text{free}|\text{spam})P(\text{spam})}{P(\text{free})}$$

$$P(\text{spam}) = \frac{30}{74}$$

$$P(\text{free}) = \frac{51}{74}$$

Bayes Classification

- In classification, the posterior probability can be interpreted as: "What is the probability that a particular example *q* belongs to class *c*, given its observed feature values q_i?"

$$P(t = c \mid q_1, \ldots, q_n)$$

- The prior probability, aka class prior, is the probability of the target feature t being the class c

$$P(t = c)$$

- Estimate the posterior probability for each class from the training data using Bayes Theorem

$$P(t = c|q_1, \ldots, q_n) = \frac{P(q_1, \ldots, q_n \mid t = c)P(t = c)}{P(q_1, \ldots, q_n)}$$

Bayes Classification

- The prediction for an example *q* is the target class *c* that has the highest posterior probability given its features values q_i

- This is known as a maximum a posteriori (MAP) prediction

$$
\begin{aligned}
\mathbb{M}_{MAP}(q) &= \underset{c \in classes(t)}{\arg\max}\; P(t = c \mid q_1, \ldots, q_n) \\
&= \underset{c \in classes(t)}{\arg\max}\; \frac{P(q_1, \ldots, q_n \mid t = c) \times P(t = c)}{P(q_1, \ldots, q_n)} \\
&= \underset{c \in classes(t)}{\arg\max}\; P(q_1, \ldots, q_n \mid t = c) \times P(t = c)
\end{aligned}
$$

Bayes Classification

- The prediction for an example *q* is the target class *c* that has the highest posterior probability given its features values q_i

- This is known as a maximum a posteriori (MAP) prediction

$$
\begin{aligned}
\mathbb{M}_{MAP}(q) &= \underset{c \in classes(t)}{\arg\max}\ P(t = c \mid q_1, \ldots, q_n) \\
&= \underset{c \in classes(t)}{\arg\max}\ \frac{P(q_1, \ldots, q_n \mid t = c) \times P(t = c)}{P(q_1, \ldots, q_n)} \\
&= \underset{c \in classes(t)}{\arg\max}\ P(q_1, \ldots, q_n \mid t = c) \times P(t = c)
\end{aligned}
$$

Aside: Conditional Independence

- Two events are said to be independent of each other if knowledge of one event has no effect on the probability of the other event
- If X and Y are independent, then

$$P(X \mid Y) = P(X)$$

$$P(X, Y) = P(X) \times P(Y)$$

- If two events X and Y, are conditionally independent given knowledge of a third event, Z, then

$$P(X \mid Y, Z) = P(X \mid Z)$$

$$P(X, Y \mid Z) = P(X \mid Z) \times P(Y \mid Z)$$

Assuming Conditional Independence

$$
\begin{aligned}
P(q_1, \ldots, q_n \mid t = c) &= P(q_1 \mid t = c) \times P(q_2 \mid t = c) \\
&\quad \times \ldots \times P(q_n \mid t = c) \\
&= \prod_{i=1}^{n} P(q_i \mid t = c)
\end{aligned}
$$

OLLSCOIL TEICNEOLAÍOCHTA BHAILE ÁTHA CLIATH
DUBLIN
TECHNOLOGICAL UNIVERSITY DUBLIN

Naïve Bayes Classification

A Naive Bayes classifier returns a MAP prediction for an example *q* where the posterior probabilities for the classes of the target feature are computed under the assumption of conditional independence

$$\mathbb{M}_{MAP}(q) = \underset{c \in classes(t)}{\arg\max} \prod_{i=1}^{n} P(q_i \mid t = c) \times P(t = c)$$

Example

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

What is the diagnosis for someone with a headache, fever but no vomiting?

$$\arg\max_{c \in classes(t)} \prod_{i=1}^{m} P(q_i \mid t = c) \times P(t = c)$$

- Two classes, *m=T* & *m=F* and three features *h, f, v*
- Need to calculate
  - 2 x prior probabilities *P(m=T)* and *P(m=F)*
  - 12 conditional probabilities

    P(h=T | m=T) P(h=F | m=T) P(h=T | m=F) P(h=F | m=F)

    P(f=T | m=T) P(f=F | m=T) P(f=T | m=F) P(f=F | m=F)

    P(v=T | m=T) P(v=F | m=T) P(v=T | m=F) P(v=F | m=F)

Example

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

What is the diagnosis for someone with a headache, fever but no vomiting?

$$\arg\max_{c \in classes(t)} \prod_{i=1}^{m} P(q_i \mid t = c) \times P(t = c)$$

Two classes: *m=T* & *m=F*

m=T: P(h=T | m=T) x P(f=T | m=T) x P(v=F | m=T) x P(m=T)

m=F: P(h=T | m=F) x P(f=T | m=F) x P(v=F | m=F) x P(m=F)

Example: Fraud detection on loan applications

ID	Credit History	Guarantor/ CoApplicant	Accomodation	Fraud
1	current	none	own	true
2	paid	none	own	false
3	paid	none	own	false
4	paid	guarantor	rent	true
5	arrears	none	own	false
6	arrears	none	own	true
7	current	none	own	false
8	arrears	none	own	false
9	current	none	rent	false
10	none	none	own	true
11	current	coapplicant	own	false
12	current	none	own	true
13	current	none	rent	true
14	paid	none	own	false
15	arrears	none	own	false
16	current	none	own	false
17	arrears	coapplicant	rent	false
18	arrears	none	free	false
19	arrears	none	own	false
20	paid	none	own	false

How many prior probabilities?

How many conditional probabilities?

What is the prediction for someone with credit history paid, a guarantor, and free accommodation?

fr: P(CH=paid | fr) x P(GC=guarantor | fr) x P(ACC=free | fr) x P(fr)

¬fr: P(CH=paid |¬fr) x P(GC=guarantor |¬fr) x P(ACC=free |¬fr) x P(¬f)

Example: Fraud detection on loan applications

fr: P(CH=paid | fr) x P(GC=guarantor | fr) x P(ACC=free | fr) x P(fr)

¬fr: P(CH=paid |¬fr) x P(GC=guarantor |¬fr) x P(ACC=free |¬fr) x P(¬f)

$$
\begin{array}{rclrcl}
P(fr) & = & 6/20 & P(\neg fr) & = & 14/20 \\
P(CH = paid \mid fr) & = & 1/6 & P(CH = paid \mid \neg fr) & = & 4/14 \\
P(GC = guarantor \mid fr) & = & 1/6 & P(GC = guarantor \mid \neg fr) & = & 0/14 \\
P(ACC = free \mid fr) & = & 0/6 & P(ACC = free \mid \neg fr) & = & 1/14
\end{array}
$$

$$\left(\prod_{k=1}^{n} P(q_k \mid fr)\right) \times P(fr) = 0.0$$

$$\left(\prod_{k=1}^{n} P(q_k \mid \neg fr)\right) \times P(\neg fr) = 0.0$$

Smoothing

- Smoothing takes some of the probability from the events with lots of the probability share and gives it to the other probabilities in the set
- Many different ways to smooth probabilities
- Laplace smoothing (conditional probabilities)

$$P(f = v|c) = \frac{count(f = v|c) + k}{count(f|c) + (k \times |Domain(f)|)}$$

where

count(f=v | c) is how often the feature *f* has value *v* for instances where the class is *c*

count(f | c) is how often the feature *f* has any value where the class is *c*

|Domain(f)| is the number of different values feature *f* can have

k is a parameter (normally 1, 2 or 3)

Smoothing

Apply Laplace Smoothing to GC feature for Not Fraud class

Raw	$P(GC = none \mid \neg fr)$	$=$	0.8571
Probabilities	$P(GC = guarantor \mid \neg fr)$	$=$	0
	$P(GC = coapplicant \mid \neg fr)$	$=$	0.1429
Smoothing	k	$=$	3
Parameters	$count(GC \mid \neg fr)$	$=$	14
	$count(GC = none \mid \neg fr)$	$=$	12
	$count(GC = guarantor \mid \neg fr)$	$=$	0
	$count(GC = coapplicant \mid \neg fr)$	$=$	2
	$\lvert Domain(GC) \rvert$	$=$	3
Smoothed	$P(GC = none \mid \neg fr) = \frac{12+3}{14+(3\times3)}$	$=$	0.6522
Probabilities	$P(GC = guarantor \mid \neg fr) = \frac{0+3}{14+(3\times3)}$	$=$	0.1304
	$P(GC = coapplicant \mid \neg fr) = \frac{2+3}{14+(3\times3)}$	$=$	0.2174

Laplace Smoothing

What is the prediction for someone with credit history paid, a guarantor, and free accommodation?

Smoothed probabilities

$$\begin{aligned} P(fr) &= 0.3 & P(\neg fr) &= 0.7 \\ P(CH = paid \mid fr) &= 0.2222 & P(CH = paid| \mid \neg fr) &= 0.2692 \\ P(GC = guarantor \mid fr) &= 0.2667 & P(GC = guarantor \mid \neg fr) &= 0.1304 \\ P(ACC = free \mid fr) &= 0.2 & P(ACC = free \mid \neg fr) &= 0.1739 \end{aligned}$$

$$\left(\prod_{k=1}^{n} P(q_k \mid fr)\right) \times P(fr) = 0.0036$$

$$\left(\prod_{k=1}^{n} P(q_k \mid \neg fr)\right) \times P(\neg fr) = 0.0043$$

Raw probabilities

$$\begin{aligned} P(fr) &= 0.3 & P(\neg fr) &= 0.7 \\ P(CH = paid \mid fr) &= 0.1666 & P(CH = paid \mid \neg fr) &= 0.2857 \\ P(GC = guarantor \mid fr) &= 0.1666 & P(GC = guarantor \mid \neg fr) &= 0 \\ P(ACC = free \mid fr) &= 0 & P(ACC = free \mid \neg fr) &= 0.0714 \end{aligned}$$

$$\left(\prod_{k=1}^{n} P(q_k \mid fr)\right) \times P(fr) = 0.0$$

$$\left(\prod_{k=1}^{n} P(q_k \mid \neg fr)\right) \times P(\neg fr) = 0.0$$

Continuous Features

Two ways to handle continuous features:

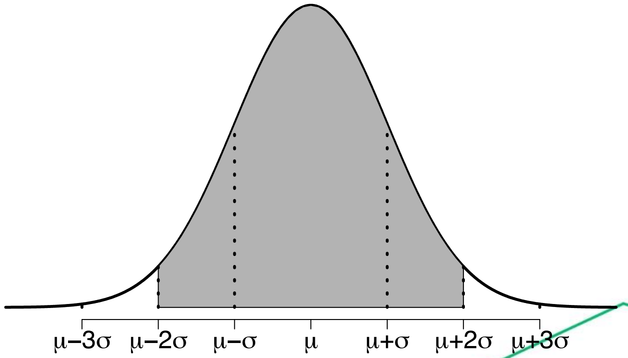
- Use a Probability Density Function (PDF)
    - fit the most appropriate PDF to the data and using it to calculate the conditional probabilities for the test instance
- Use Binning
    - convert the feature to a categorical feature using binning

Using PDFs

- A probability density function (PDF) represents the probability distribution of a continuous feature using a mathematical function
  - e.g. normal distribution

$$N(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

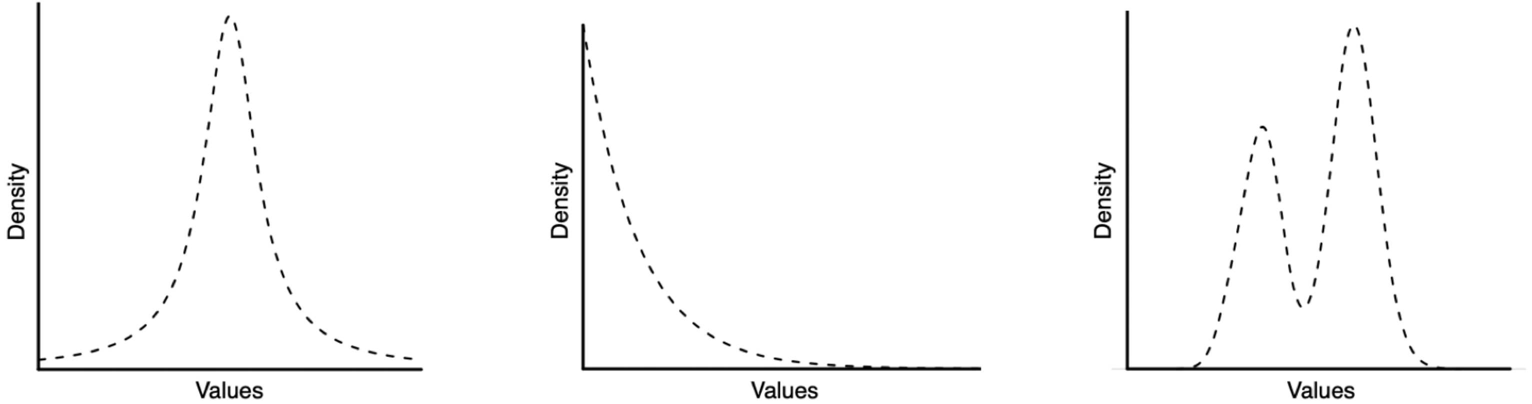
- A PDF defines a density curve, the shape is determined by
  1. the statistical distribution used to define the PDF
  2. the values of the parameters e.g. μ σ for normal dist



Definitions of some standard PDFs

Normal $x \in \mathbb{R}$ $\mu \in \mathbb{R}$ $\sigma \in \mathbb{R}_{>0}$	$N(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
Student-*t* $x \in \mathbb{R}$ $\phi \in \mathbb{R}$ $\rho \in \mathbb{R}_{>0}$ $\kappa \in \mathbb{R}_{>0}$ $z = \frac{x-\phi}{\rho}$	$\tau(x, \phi, \rho, \kappa) = \frac{\Gamma(\frac{\kappa+1}{2})}{\Gamma(\frac{\kappa}{2}) \times \sqrt{\pi\kappa} \times \rho} \times \left(1 + \left(\frac{1}{\kappa} \times z^2\right)\right)^{-\frac{\kappa+1}{2}}$
Exponential $x \in \mathbb{R}$ $\lambda \in \mathbb{R}_{>0}$	$E(x, \lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$
Mixture of *n* Gaussians $x \in \mathbb{R}$ $\{\mu_1, \ldots, \mu_n \mid \mu_i \in \mathbb{R}\}$ $\{\sigma_1, \ldots, \sigma_n \mid \sigma_i \in \mathbb{R}_{>0}\}$ $\{\omega_1, \ldots, \omega_n \mid \omega_i \in \mathbb{R}_{>0}\}$ $\sum_{i=1}^{n} \omega_i = 0$	$N(x, \mu_1, \sigma_1, \omega_1, \ldots, \mu_n, \sigma_n, \omega_n) = \sum_{i=1}^{n} \frac{\omega_i}{\sigma_i\sqrt{2\pi}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}$

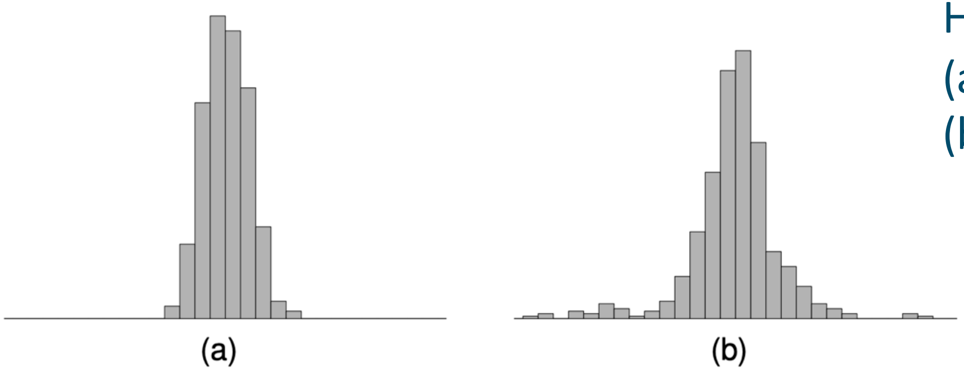
Plots of some standard PDFs



(a) Normal/Student-t (b) Exponential (c) Mixture of Gaussians

student-t distribution is more robust to outliers than the normal distribution

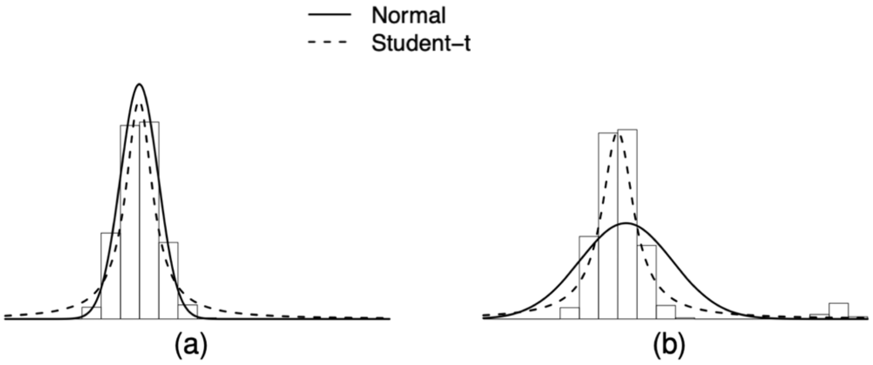
Student-t vs Normal Distribution



Histograms of two datasets:

(a) has light tails

(b) has fat tails, more outliers



Overlaid with PDFs of student-t and normal distributions that have been fitted to the data shows that student-t is less affected by the outliers

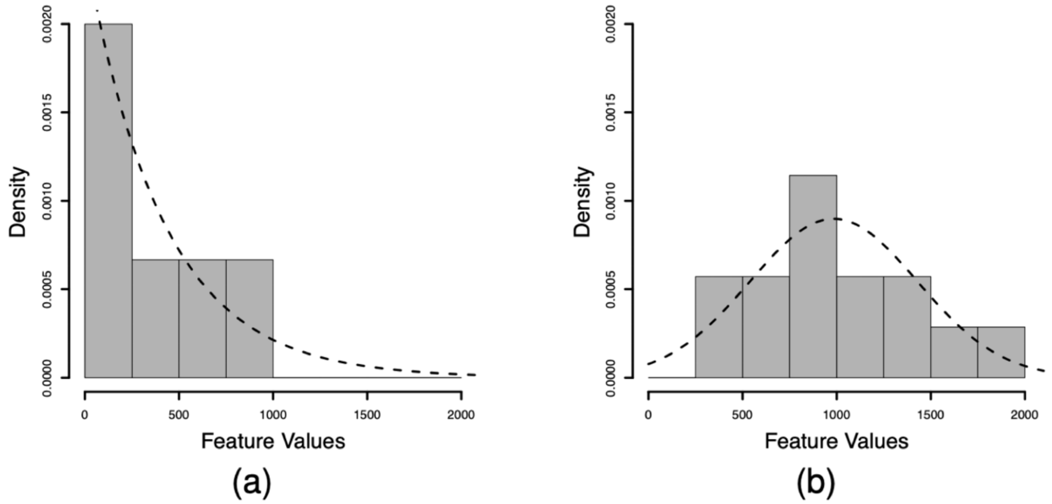
Using Continuous Features

ID	Credit History	Guarantor/ CoApplicant	Accommodation	Account Balance	Fraud
1	current	none	own	56.75	true
2	current	none	own	1,800.11	false
3	current	none	own	1,341.03	false
4	paid	guarantor	rent	749.50	true
5	arrears	none	own	1,150.00	false
6	arrears	none	own	928.30	true
7	current	none	own	250.90	false
8	arrears	none	own	806.15	false
9	current	none	rent	1,209.02	false
10	none	none	own	405.72	true
11	current	coapplicant	own	550.00	false
12	current	none	free	223.89	true
13	current	none	rent	103.23	true
14	paid	none	own	758.22	false
15	arrears	none	own	430.79	false
16	current	none	own	675.11	false
17	arrears	coapplicant	rent	1,657.20	false
18	arrears	none	free	1,405.18	false
19	arrears	none	own	760.51	false
20	current	none	own	985.41	false

Define two PDFs for the new feature conditional probabilities, PDFs do not have to have the same statistical distribution

$$P(AB = x \mid \mathrm{fr}) = PDF_1(AB = x \mid \mathrm{fr})$$
$$P(AB = x \mid \neg\mathrm{fr}) = PDF_2(AB = x \mid \neg\mathrm{fr})$$

Step 1: Select the appropriate dist



Histograms (bin size of 250) of the AC feature
(a) fraud instances overlaid with an exponential distribution
(b) non fraud instances overlaid with a normal distribution

Step 2: Fix distribution parametres

$$P(AB = x \mid \text{fr}) = PDF_1(AB = x \mid \text{fr})$$

Exponential dist:
λ estimated as 1 / sample mean

$$\lambda = 1/\bar{x}$$

ID	...	Account Balance	Fraud
1		56.75	true
4		749.50	true
6		928.30	true
10	...	405.72	true
12		223.89	true
13		103.23	true
$\overline{AB}$		411.22	
$\lambda = {}^{1}/_{\overline{AB}}$		0.0024	

Step 2: Fix distribution parametres

$$P(AB = x \mid \neg\mathrm{fr}) = PDF_2(AB = x \mid \neg\mathrm{fr})$$

Normal dist:
μ estimated as sample mean
σ estimated as sample st dev

ID	...	Account Balance	Fraud
2		1 800.11	false
3		1 341.03	false
5		1 150.00	false
7		250.90	false
8		806.15	false
9		1 209.02	false
11		550.00	false
14		758.22	false
15		430.79	false
16		675.11	false
17		1 657.20	false
18		1 405.18	false
19		760.51	false
20		985.41	false
$\overline{AB}$		984.26	
stdev(AB)		460.94	

SCOIL TEICNEOLAÍOCHTA ...ILE ÁTHA CLIATH ...BLIN TECHNOLOGICAL UNIVERSITY DUBLIN

Example

What is the prediction for someone with credit history paid, a guarantor, free accommodation and an account balance of 759.07?

ID	Credit History	Guarantor/ CoApplicant	Accommodation	Account Balance	Fraud
1	current	none	own	56.75	true
2	current	none	own	1,800.11	false
3	current	none	own	1,341.03	false
4	paid	guarantor	rent	749.50	true
5	arrears				
6	arrears				
7	current				
8	arrears				
9	current				
10	none				
11	current				
12	current				
13	current				
14	paid				
15	arrears				
16	current				
17	arrears				
18	arrears				
19	arrears				
20	current				

Smoothed probabilities

$$P(fr) = 0.3 \qquad P(\neg fr) = 0.7$$

$$P(CH = paid | fr) = 0.2222 \qquad P(CH = paid | \neg fr) = 0.2692$$

$$P(GC = guarantor | fr) = 0.2667 \qquad P(GC = guarantor | \neg fr) = 0.1304$$

$$P(ACC = free | fr) = 0.2 \qquad P(ACC = free | \neg fr) = 0.1739$$

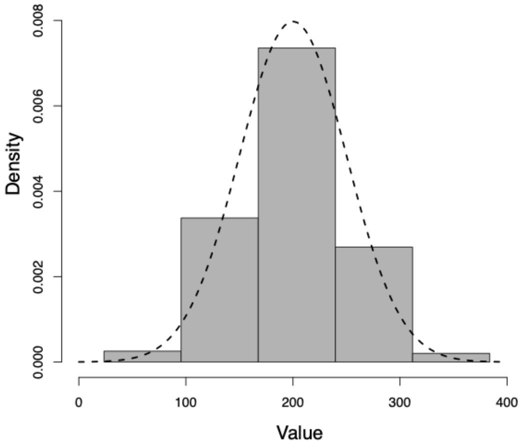
$$P(AB = 759.07 | fr) \approx E\left(\begin{matrix}759.07,\\ \lambda = 0.0024\end{matrix}\right) = 0.00039 \qquad P(AB = 759.07 | \neg fr) \approx N\left(\begin{matrix}759.07,\\ \mu = 984.26,\\ \sigma = 460.94\end{matrix}\right) = 0.00077$$

$$\left(\prod_{k=1}^{n} P(q_k | fr)\right) \times P(fr) = 0.0000014$$

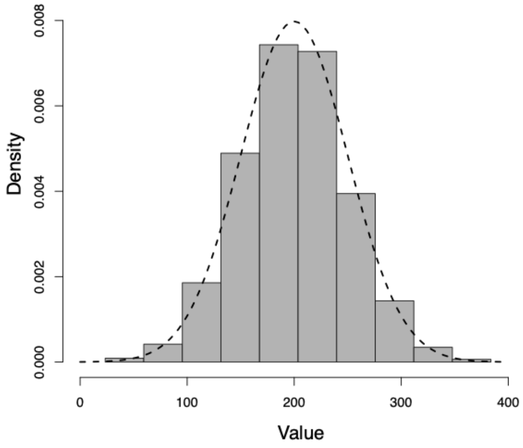
$$\left(\prod_{k=1}^{n} P(q_k | \neg fr)\right) \times P(\neg fr) = 0.0000033$$

Binning

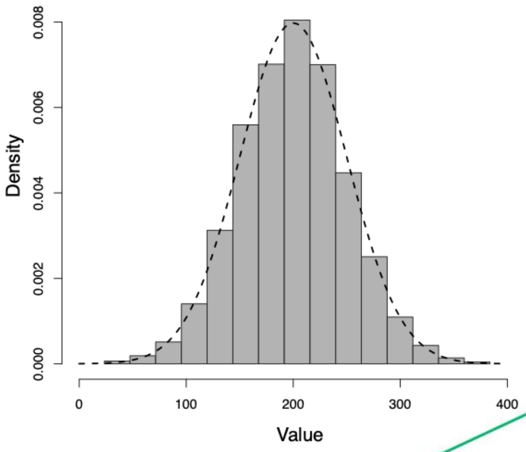
Equal-width binning splits the range of the feature values into *b* bins each of size *range/b*



5 equal-width bins



10 equal-width bins

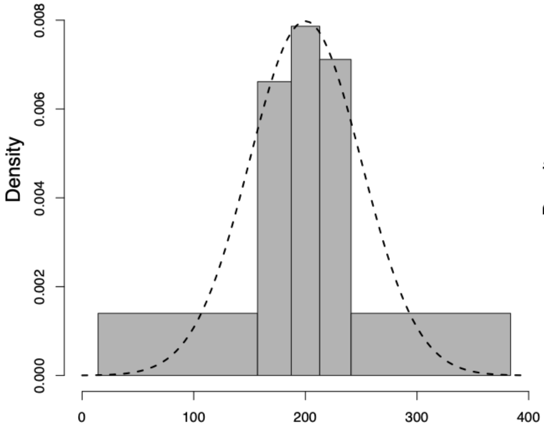


15 equal-width bins

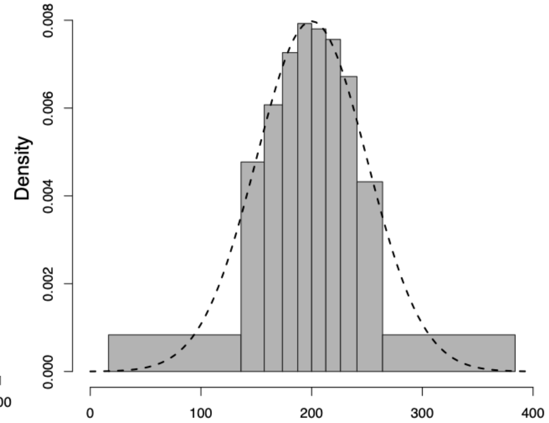
OLLSCOIL TEICNEOLAÍOCHTA BHAILE ÁTHA CLIATH
DUBLIN
TECHNOLOGICAL UNIVERSITY DUBLIN

Binning

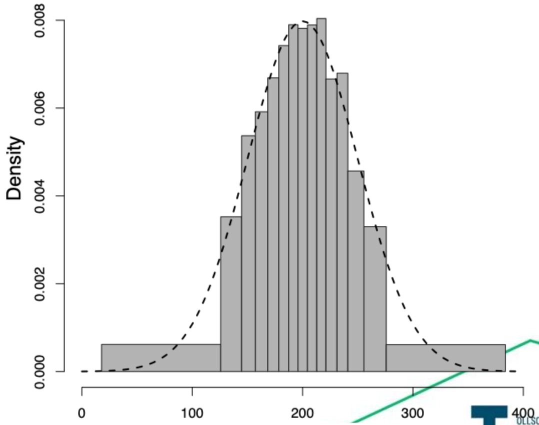
- Equal-frequency binning sorts the feature values into ascending order and places an equal number of instances into each bin
- The number of instances placed in each bin is the total number of instances divided by the number of bins, *b*



5 equal-freq bins



10 equal-freq bins



15 equal-freq bins

Equal-frequency Binning is recommended

Example - Training

ID	Credit History	Guarantor/ CoApplicant	Accommodation	Account Balance	Loan Amount	Fraud
1	current	none	own	56.75	900	true
2	current	none	own	1 800.11	150 000	false
3	current	none	own	1 341.03	48 000	false
4	paid	guarantor	rent	749.50	10 000	true
5	arrears	none	own	1 150.00	32 000	false
6	arrears	none	own	928.30	250 000	true
7	current	none	own	250.90	25 000	false
8	arrears	none	own	806.15	18 500	false
9	current	none	rent	1 209.02	20 000	false
10	none	none	own	405.72	9 500	true
11	current	coapplicant	own	550.00	16 750	false
12	current	none	free	223.89	9 850	true
13	current	none	rent	103.23	95 500	true
14	paid	none	own	758.22	65 000	false
15	arrears	none	own	430.79	500	false
16	current	none	own	675.11	16 000	false
17	arrears	coapplicant	rent	1 657.20	15 450	false
18	arrears	none	free	1 405.18	50 000	false
19	arrears	none	own	760.51	500	false
20	current	none	own	985.41	35 000	false

Example - Training

Loan amount feature discretised into 4 equal-frequency bins

ID	Loan Amount	Binned Loan Amount	Fraud
15	500	bin1	false
19	500	bin1	false
1	900	bin1	true
10	9,500	bin1	true
12	9,850	bin1	true
4	10,000	bin2	true
17	15,450	bin2	false
16	16,000	bin2	false
11	16,750	bin2	false
8	18,500	bin2	false

ID	Loan Amount	Binned Loan Amount	Fraud
9	20,000	bin3	false
7	25,000	bin3	false
5	32,000	bin3	false
20	35,000	bin3	false
3	48,000	bin3	false
18	50,000	bin4	false
14	65,000	bin4	false
13	95,500	bin4	true
2	150,000	bin4	false
6	250,000	bin4	true

Example - Training

Calculate conditional probabilities

ID	Loan Amount	Binned Loan Amount	Fraud
15	500	bin1	false
19	500	bin1	false
1	900	bin1	true
10	9,500	bin1	true
12	9,850	bin1	true
4	10,000	bin2	true
17	15,450	bin2	false
16	16,000	bin2	false
11	16,750	bin2	false
8	18,500	bin2	false

ID	Loan Amount	Binned Loan Amount	Fraud
9	20,000	bin3	false
7	25,000	bin3	false
5	32,000	bin3	false
20	35,000	bin3	false
3	48,000	bin3	false
18	50,000	bin4	false
14	65,000	bin4	false
13	95,500	bin4	true
2	150,000	bin4	false
6	250,000	bin4	true

$$P(\text{BLA} = bin1 \mid fr) = 3/6$$
$$P(\text{BLA} = bin1 \mid \neg fr) = 2/14$$
$$\ldots$$
$$P(\text{BLA} = bin3 \mid fr) = 0/6$$
$$P(\text{BLA} = bin3 \mid \neg fr) = 5/14$$

Need Smoothing

OLLSCOIL TEICNEOLAÍOCHTA BHAILE ÁTHA CLIATH
DUBLIN
TECHNOLOGICAL UNIVERSITY DUBLIN

Example - Training

Use Laplace smoothing with *k = 3*

ID	Loan Amount	Binned Loan Amount	Fraud
15	500	bin1	false
19	500	bin1	false
1	900	bin1	true
10	9,500	bin1	true
12	9,850	bin1	true
4	10,000	bin2	true
17	15,450	bin2	false
16	16,000	bin2	false
11	16,750	bin2	false
8	18,500	bin2	false

ID	Loan Amount	Binned Loan Amount	Fraud
9	20,000	bin3	false
7	25,000	bin3	false
5	32,000	bin3	false
20	35,000	bin3	false
3	48,000	bin3	false
18	50,000	bin4	false
14	65,000	bin4	false
13	95,500	bin4	true
2	150,000	bin4	false
6	250,000	bin4	true

$$P(\text{BLA} = bin1 \mid fr) = 3/6$$
$$P(\text{BLA} = bin1 \mid \neg fr) = 2/14$$
$$\ldots$$
$$P(\text{BLA} = bin3 \mid fr) = 0/6$$
$$P(\text{BLA} = bin3 \mid \neg fr) = 5/14$$

$$P(f = v|c) = \frac{count(f = v|c) + k}{count(f|c) + (k \times |Domain(f)|)}$$

$$P(\text{BLA} = bin3 \mid fr) = \frac{0+3}{6+(3\times 4)} = 0.166$$
$$P(\text{BLA} = bin3 \mid \neg fr) = \frac{5+3}{14+(3\times 4)} = 0.307$$

Example - Testing

Identify bin thresholds

ID	Loan Amount	Binned Loan Amount	Fraud
15	500	bin1	false
19	500	bin1	false
1	900	bin1	true
10	9,500	bin1	true
12	9,850	bin1	true
4	10,000	bin2	true
17	15,450	bin2	false
16	16,000	bin2	false
11	16,750	bin2	false
8	18,500	bin2	false

9,925

19,250

ID	Loan Amount	Binned Loan Amount	Fraud
9	20,000	bin3	false
7	25,000	bin3	false
5	32,000	bin3	false
20	35,000	bin3	false
3	48,000	bin3	false
18	50,000	bin4	false
14	65,000	bin4	false
13	95,500	bin4	true
2	150,000	bin4	false
6	250,000	bin4	true

49,000

Bin Thresholds		
	Bin1	$\leq 9,925$
$9,925 <$	Bin2	$\leq 19,250$
$19,250 <$	Bin3	$\leq 49,000$
$49,000 <$	Bin4	

Bin thresholds are needed to determine the appropriate bin for each query instance before making a prediction for it

Summary

- Naive Bayes, the most common probabilistic approach to prediction, is an *eager based* learning approach based on Bayes Theorem
- Robust as the accuracy of the conditional probabilities do not necessarily translate to prediction errors
  - Concerned with the relative values of the conditional probabilities for the target classes rather than the exact probabilities -> not good for predicting continuous targets
- Robust to the curse of dimensionality due to the assumption of conditional independence -> but cannot handle interactions between features
- Can handle missing values by dropping conditional probabilities for features taking values not in the data -> good on sparse datasets (text)

Remember

- Naive Bayes prediction relies on the assumption that all the features are conditionally independent, i.e. the value of any feature is unrelated to the presence or absence of any feature given the class label
- In some domains violations of the independence assumption can lead to poor performance by Naive Bayes

- Always need to keep in mind the type of data and the type of problem to be solved when choosing a prediction algorithm

OLLSCOIL TEICNEOLAÍOCHTA BHAILE ÁTHA CLIATH
TU DUBLIN
TECHNOLOGICAL UNIVERSITY DUBLIN

Naïve Bayes in scikit-learn

NB classifiers in sklearn differ by the assumptions they make regarding $P(x_i|t)$

CategoricalNB: similar to what is described in the lecture, i.e. categorical features

GaussianNB: the likelihood of the features is assumed to be Gaussian, i.e. continuous features

BernoulliNB: data is distributed according to multivariate Bernoulli distributions, i.e. many features, each one is binary

MultinomialNB: a variant of Categorical NB good for text data

ComplementNB: a variant of MNB that handles imbalanced data

Questions?