Business Statistics 41000

Lecture 4

Mladen Kolar

Summarizing dependence

Joint distributions provide us with an additional ability to characterize relationships between two measurements.

Comprehensive information is captured by the conditional distributions themselves. But, as with the mean, median, mode and variance, we may want a more **parsimonious** description of the relationship between two quantities.

Standardizing

In order to summarize the relationship between two random variables X and Y we begin by **standardizing** them so they share a common scale.

We do this by defining two new random variables

$$Z_X = \frac{X - E(X)}{\sqrt{\mathrm{Var}(X)}}$$

and

$$Z_Y = \frac{Y - E(Y)}{\sqrt{\mathrm{Var}(Y)}}.$$

It is a straightforward calculation to verify that $E(Z_X) = E(Z_Y) = 0$ and $\mathrm{Var}(Z_X) = \mathrm{Var}(Z_Y) = 1$.

Correlation for Random Variables

The **correlation** between two random variables X and Y is

$$\rho_{XY} = \text{cor}(X, Y) = E(Z_X Z_Y),$$

where Z_X and Z_Y are standardized versions of X and Y.

Expectation of a product of two discrete random variables

$$E(Z_X Z_Y) = \sum_{z_X, z_Y} z_X z_Y P(Z_X = z_X, Z_Y = z_Y).$$

Correlation is the expectation of the product of two standardized random variables.

Covariance for Random Variables

The covariance between two discrete random variables X and Y is given by

$$\sigma_{XY} = \text{cov}(X, Y) = \rho_{XY} \cdot \sigma_X \cdot \sigma_Y,$$

where σ_X, σ_Y are standard deviations of X and Y.

Alternative way to compute the covariance

$$\begin{aligned}\sigma_{XY} &= \text{cov}(X, Y) = E[(X - E(X))(Y - E(Y))] \\ &= E(XY) - E(X)E(Y)\end{aligned}$$

Calculating $\mathrm{cov}(X, Y)$

		Y			
		1	2	3	
X	0	1/8	0	0	1/8
	1	0	1/4	1/8	3/8
	2	0	1/4	1/8	3/8
	3	1/8	0	0	1/8
		1/4	1/2	1/4	

$$E[X] = 3/8 + 2 \times 3/8 + 3 \times 1/8 = 3/2$$

$$E[Y] = 1/4 + 2 \times 1/2 + 3 \times 1/4 = 2$$

Calculating $\mathrm{cov}(X, Y)$

		Y			
		1	2	3	
X	0	1/8	0	0	1/8
	1	0	1/4	1/8	3/8
	2	0	1/4	1/8	3/8
	3	1/8	0	0	1/8
		1/4	1/2	1/4	

$$
\begin{aligned}
E[X] &= 3/8 + 2 \times 3/8 + 3 \times 1/8 = 3/2 \\
E[Y] &= 1/4 + 2 \times 1/2 + 3 \times 1/4 = 2 \\
E[XY] &= 1/4 \times (2+4) + 1/8 \times (3+6+3) \\
&= 3
\end{aligned}
$$

Calculating $\mathrm{cov}(X, Y)$

		Y			
		1	2	3	
X	0	1/8	0	0	1/8
	1	0	1/4	1/8	3/8
	2	0	1/4	1/8	3/8
	3	1/8	0	0	1/8
		1/4	1/2	1/4	

$$
\begin{aligned}
E[X] &= 3/8 + 2 \times 3/8 + 3 \times 1/8 = 3/2 \\
E[Y] &= 1/4 + 2 \times 1/2 + 3 \times 1/4 = 2 \\
E[XY] &= 1/4 \times (2+4) + 1/8 \times (3+6+3) \\
&= 3 \\
\sigma_{XY} &= E[XY] - E[X]E[Y]
\end{aligned}
$$

Calculating cov(X, Y)

		Y			
		1	2	3	
	0	$1/8$	0	0	$1/8$
X	1	0	$1/4$	$1/8$	$3/8$
	2	0	$1/4$	$1/8$	$3/8$
	3	$1/8$	0	0	$1/8$
		$1/4$	$1/2$	$1/4$	

$$
\begin{aligned}
E[X] &= 3/8 + 2 \times 3/8 + 3 \times 1/8 = 3/2 \\
E[Y] &= 1/4 + 2 \times 1/2 + 3 \times 1/4 = 2 \\
E[XY] &= 1/4 \times (2+4) + 1/8 \times (3+6+3) \\
&= 3 \\
\sigma_{XY} &= E[XY] - E[X]E[Y] \\
&= 3 - 3/2 \times 2 = 0
\end{aligned}
$$

Calculating $\mathrm{cov}(X, Y)$

		Y			
		1	2	3	
X	0	1/8	0	0	1/8
	1	0	1/4	1/8	3/8
	2	0	1/4	1/8	3/8
	3	1/8	0	0	1/8
		1/4	1/2	1/4	

$$
\begin{aligned}
E[X] &= 3/8 + 2 \times 3/8 + 3 \times 1/8 = 3/2 \\
E[Y] &= 1/4 + 2 \times 1/2 + 3 \times 1/4 = 2 \\
E[XY] &= 1/4 \times (2+4) + 1/8 \times (3+6+3) \\
&= 3 \\
\sigma_{XY} &= E[XY] - E[X]E[Y] \\
&= 3 - 3/2 \times 2 = 0 \\
\rho_{XY} &= \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = 0
\end{aligned}
$$

Zero Covariance versus Independence

From this example we learn that zero covariance (correlation) **does not** imply independence.

However, it turns out that independence **does** imply zero covariance (correlation).

Mean and Variance of a Linear Combination

Expected Value of Sum = Sum of Expected Values

$$E[c_1X_1 + c_2X_2 + \ldots + c_nX_n] = c_1E[X_1] + c_2E[X_2] + \ldots + c_nE[X_n]$$

regardless of how the r.v.s $X_1, \ldots, X_n$ are related to each other.

In particular it doesn't matter if they're dependent or independent.

Variance of a Sum $\neq$ Sum of Variances!

$$\begin{aligned} Var(aX + bY) &= a^2 Var(X) + b^2 Var(Y) + 2ab\text{cov}(X, Y) \\ &= a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho_{XY}\sigma_X\sigma_Y \end{aligned}$$

Independence $\Rightarrow Var(X + Y) = Var(X) + Var(Y)$

X and Y independent $\implies \text{cov}(X, Y) = 0$. Hence, independence implies

$$\begin{aligned} Var(X + Y) &= Var(X) + Var(Y) + 2\text{cov}(X, Y) \\ &= Var(X) + Var(Y) \end{aligned}$$

Also true for three or more RVs

If $X_1, X_2, \ldots, X_n$ are independent, then

$$Var(X_1 + X_2 + \ldots X_n) = Var(X_1) + Var(X_2) + \ldots + Var(X_n)$$

Naïve Bayes Classification

An accounting company needs to audit annual financial reports submitted by large companies. The result of an audit can be fraudulent ($Y = 1$) or truthful ($Y = 0$). The accounting firm has strong incentives to accurately identify fraudulent reports.

The accounting firm also has information on whether or not the customer has had prior legal trouble (criminal or civil charges of any kind).

- Let $X = 1$ denote prior legal trouble, and
- $X = 0$ if no prior legal trouble.

This information has not been used in previous audits, but potentially has some value.

The accounting company has information of 1000 companies it investigated in the past.

		X		
		Prior Legal	No Prior Legal	
		1	0	Total
Y	1 (fraudulent)	50	50	100
	0 (truthful)	180	720	900
	Total	230	770	1000

The new company had prior legal trouble.
What is the probability it belongs to fraudulent class?

$$P(Y=1 \mid X=1) = \frac{P(X=1 \mid Y=1)P(Y=1)}{P(X=1)} = \frac{\frac{50}{100}\frac{100}{1000}}{\frac{230}{1000}} = \frac{50}{230}$$

$$P(Y=0 \mid X=1) = \frac{P(X=1 \mid Y=0)P(Y=0)}{P(X=1)} = \frac{\frac{180}{900}\frac{900}{1000}}{\frac{230}{1000}} = \frac{180}{230}$$

Bayes Rule

$$P(Y = y \mid X = x) = \frac{P(X = x \mid Y = y)P(Y = y)}{P(X = x)}$$

Common terminology:

- $P(Y)$ - prior
- $P(X \mid Y)$ - likelihood
- $P(X)$ - normalization

In order to decide whether the financial report is fraudulent or not, we simply compare $P(Y = 1 \mid X = x)$ to $P(Y = 0 \mid X = x)$

If

$$P(Y = 1 \mid X = x) > P(Y = 0 \mid X = x)$$

then we classify the report as fradulent, otherwise we classify it as truthful.

Notice that using the Bayes rule, we simply compare if

$$P(X = x \mid Y = 1)P(Y = 1) > P(X = x \mid Y = 0)P(Y = 0)$$

and we do not need to compute the normalization $P(X = x)$.

The goal of classification is to learn a mapping from input variables to the target class.

- Classifier: $f : X \mapsto Y$
- X are input variables
- Y is the target class

Examples:

- Spam detection
- Language identification
- Sentiment analysis
  - Positive/negative movie or product review
- Suggesting tags on a question answering forum

Example: A dataset from a loan application fraud detection domain.

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

Credit History	Guarantor/CoApplicant	Accomodation	Fraudulent
paid	none	rent	?

Random variable F represents whether the loan is fraudulent or not.

Random variable CH represents `credit history`.

- It takes values: `none`, `paid`, `current`, `arrears`

Random variable GC represents `Guarantor/CoApplicant`.

- It takes values: `none`, `coapplicant`, `guarantor`.

Random variable A represents `Accomodatoin`.

- It takes values: `own`, `rent`, `free`.

Recall that in order to compare $P(F = t \mid CH, GC, A)$ with $P(F = f \mid CH, GC, A)$ we need

- prior, $P(F = t)$
- likelihoods, $P(CH, GC, A \mid F = t)$ and $P(CH, GC, A \mid F = f)$

Where do these probabilities come from?

How many different probabilities do we need?

Prior, $P(F)$:

- we need 1 parameter: $P(F = t)$
- recall, $P(F = f) = 1 - P(F = t)$

Likelihood, $P(CH, GC, A \mid F)$

- total number of parameters $2 \cdot (4 \cdot 3 \cdot 3 - 1) = 70$
- why?

This is a large number of parameters.

We will make some assumptions and get back to this problem.

Independence of two random variables

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

Y and X do not contain information about each other.
Observing Y does not help predicting X.
Observing X does not help predicting Y.

Examples:

- Independent: Winning on roulette this week and next week.
- Dependent: Russian roulette

Conditional independence

X is **conditionally independent** of Y given Z, if for all values of (i, j, k) that random variables X, Y, and Z can take, we have

$$P(X = i, Y = j \mid Z = k) = P(X = i \mid Z = k) \cdot P(Y = j \mid Z = k)$$

Knowing Z makes X and Y independent.

Examples:

- Shoe size and reading skills are dependent. Given *age*, shoe size and reading skills are independent.
- *Storks deliver babies*. Highly statistically significant correlation exists between stork populations and human birth rates across Europe.

London taxi drivers:

A survey has pointed out a positive and significant correlation between the number of accidents and wearing coats. They concluded that coats could hinder movements of drivers and be the cause of accidents. A new law was prepared to prohibit drivers from wearing coats when driving.

Finally another study pointed out that people wear coats when it rains. . .

$$P(\text{accidents,coat} \mid \text{rain}) = P(\text{accidents} \mid \text{rain}) \cdot P(\text{coat} \mid \text{rain})$$

Conditional independence

An equivalent definition

X is **conditionally independent** of Y given Z, if for all values of (i, j, k) that random variables X, Y, and Z can take, we have

$$P(X = i \mid Y = j, Z = k) = P(X = i \mid Z = k)$$

Example:

$$P(\text{thunder} \mid \text{rain, lightning}) = P(\text{thunder} \mid \text{lightning})$$

Thunder and rain are not independent. However, if I tell you that there is lightning, they become independent.

How can we use conditional independence in classification?

Goal: Predict Thunder

Input variables are conditionally independent

- lightning
- rain

Recall: $P(T \mid L, R) \propto P(L, R \mid T) \cdot P(T)$

How many probabilities do we need to estimate?

How can we use conditional independence in classification?

How many probabilities do we need to estimate?

Without conditional independence, we need 6 parameters to represent $P(L, R \mid T)$.

However, we have $L \perp R \mid T$, so

$$P(L, R \mid T) = P(L \mid T) \cdot P(R \mid T)$$

and we need only 4 probabilities.

The Naïve Bayes assumption

Input variables are independent given class:

$$P(X_1, X_2 \mid Y) = P(X_1 \mid Y) \cdot P(X_2 \mid Y)$$

More generally, if we have p input variables:

$$P(X_1, \ldots, X_p \mid Y) = P(X_1 \mid Y) \cdot P(X_2 \mid Y) \cdot \ldots \cdot P(X_p \mid Y)$$

The likelihood is product of individual input variable likelihoods.
How many parameters do we need now?

The Naïve Bayes assumption

How many parameters for $P(X_1, \ldots, X_p \mid Y)$?

- Without assumption we need $k \cdot (2^p - 1)$ parameters

With the Naïve Bayes assumption

$$P(X_1, \ldots, X_p \mid Y) = \prod_{i=1}^{p} P(X_i \mid Y)$$

we need $p \cdot k$ parameters.

Nice reduction, however, it may be to aggressive.

The Naïve Bayes classifier

Given:

- Prior $P(Y)$
- p conditionally independent input variables X given the class Y
- For each X_i, we have likelihood $P(X_i \mid Y)$

Decision rule:

$$P(Y = 1) \cdot \prod_{i=1}^{p} P(X_i \mid Y = 1) > P(Y = 0) \cdot \prod_{i=1}^{p} P(X_i \mid Y = 0)$$

How do we estimate the probabilities?

We count! For a given dataset

$\mathrm{Count}(A = a, B = b) \equiv$ number of examples where $A = a$ and $B = b$

Prior

$$P(Y = y) = \frac{\mathrm{Count}(Y = y)}{n}$$

Likelihood

$$P(X_i = x_i \mid Y = y) = \frac{\mathrm{Count}(X_i = x_i, Y = y)}{\mathrm{Count}(Y = y)}$$

Example: A dataset from a loan application fraud detection domain.

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

```
## Fraud
## false  true
##    14     6

##         Credit History
## Fraud    arrears current none paid
##   false        6       4    0    4
##   true         1       3    1    1

##         Guarantor CoApplicant
## Fraud    coapplicant guarantor none
##   false            2         0   12
##   true             0         1    5

##         Accommodation
## Fraud    free own rent
##   false     1  11    2
##   true      0   4    2
```

$$P(F = 1) = 0.3 \qquad P(F = 0) = 0.7$$

$$P(\mathrm{CH} = \text{none} \mid F = 1) = 0.1666 \qquad P(\mathrm{CH} = \text{none} \mid F = 0) = 0$$

$$P(\mathrm{CH} = \text{paid} \mid F = 1) = 0.1666 \qquad P(\mathrm{CH} = \text{paid} \mid F = 0) = 0.2857$$

$$P(\mathrm{CH} = \text{current} \mid F = 1) = 0.5 \qquad P(\mathrm{CH} = \text{current} \mid F = 0) = 0.2857$$

$$P(\mathrm{CH} = \text{arrears} \mid F = 1) = 0.1666 \qquad P(\mathrm{CH} = \text{arrears} \mid F = 0) = 0.4286$$

$$P(\mathrm{GC} = \text{none} \mid F = 1) = 0.8334 \qquad P(\mathrm{GC} = \text{none} \mid F = 0) = 0.8571$$

$$P(\mathrm{GC} = \text{guarantor} \mid F = 1) = 0.1666 \qquad P(\mathrm{GC} = \text{guarantor} \mid F = 0) = 0$$

$$P(\mathrm{GC} = \text{coapplicant} \mid F = 1) = 0 \qquad P(\mathrm{GC} = \text{coapplicant} \mid F = 0) = 0.1429$$

$$P(\mathrm{A} = \text{own} \mid F = 1) = 0.6666 \qquad P(\mathrm{A} = \text{own} \mid F = 0) = 0.7857$$

$$P(\mathrm{A} = \text{rent} \mid F = 1) = 0.3333 \qquad P(\mathrm{A} = \text{rent} \mid F = 0) = 0.1429$$

$$P(\mathrm{A} = \text{free} \mid F = 1) = 0 \qquad P(\mathrm{A} = \text{free} \mid F = 0) = 0.0714$$

Credit History	Guarantor/CoApplicant	Accomodation	Fraudulent
paid	none	rent	?

$$
\begin{aligned}
P(F = 1) &= 0.3 & P(F = 0) &= 0.7 \\
P(\mathrm{CH} = \text{paid} \mid F = 1) &= 0.1666 & P(\mathrm{CH} = \text{paid} \mid F = 0) &= 0.2857 \\
P(\mathrm{GC} = \text{none} \mid F = 1) &= 0.8334 & P(\mathrm{GC} = \text{none} \mid F = 0) &= 0.8571 \\
P(\mathrm{A} = \text{rent} \mid F = 1) &= 0.3333 & P(\mathrm{A} = \text{rent} \mid F = 0) &= 0.1429
\end{aligned}
$$

$$P(\mathrm{CH} = \text{paid} \mid F = 1) \cdot P(\mathrm{GC} = \text{none} \mid F = 1) \cdot P(\mathrm{A} = \text{rent} \mid F = 1) \cdot P(F = 1) = 0.0139$$

$$P(\mathrm{CH} = \text{paid} \mid F = 0) \cdot P(\mathrm{GC} = \text{none} \mid F = 0) \cdot P(\mathrm{A} = \text{rent} \mid F = 0) \cdot P(F = 0) = 0.0245$$

Credit History	Guarantor/CoApplicant	Accomodation	Fraudulent
paid	none	rent	?

$$
\begin{aligned}
P(F=1) &= 0.3 & P(F=0) &= 0.7 \\
P(\mathrm{CH}=\text{paid} \mid F=1) &= 0.1666 & P(\mathrm{CH}=\text{paid} \mid F=0) &= 0.2857 \\
P(\mathrm{GC}=\text{none} \mid F=1) &= 0.8334 & P(\mathrm{GC}=\text{none} \mid F=0) &= 0.8571 \\
P(\mathrm{A}=\text{rent} \mid F=1) &= 0.3333 & P(\mathrm{A}=\text{rent} \mid F=0) &= 0.1429
\end{aligned}
$$

$$P(\mathrm{CH}=\text{paid} \mid F=1) \cdot P(\mathrm{GC}=\text{none} \mid F=1) \cdot P(\mathrm{A}=\text{rent} \mid F=1) \cdot P(F=1) = 0.0139$$

$$P(\mathrm{CH}=\text{paid} \mid F=0) \cdot P(\mathrm{GC}=\text{none} \mid F=0) \cdot P(\mathrm{A}=\text{rent} \mid F=0) \cdot P(F=0) = 0.0245$$

Credit History	Guarantor/CoApplicant	Accomodation	Fraudulent
paid	none	rent	false

The model is generalizing beyond the dataset!

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

Credit History	Guarantor/CoApplicant	Accomodation	Fraudulent
paid	none	rent	false

Subtleties of NB

Usually (always), features are not conditionally independent.

$$P(X_1, \ldots, X_p \mid Y) \neq \prod_{i=1}^{p} P(X_i \mid Y)$$

Actual probabilities $P(Y \mid X)$ often biased towards 0 or 1.

Nonetheless, NB is the single most used classifier out there. NB often performs well, even when the assumption is violated.

Text classification

- classify e-mails (spam, ham)
- classify news articles (what is the topic of the article)
- classify reviews (positive or negative review)

Features X are entire documents (reviews):

> *I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet.*

NB classification is a popular tool for classifying text.

To battle spam, back in the day, people would carefully craft rules:

- For example: If email "FROM:black-list-address" OR (contains terms "dollars" AND "have been selected") then "SPAM"

These rules are hard to craft and experts cost a lot of money. Spammers are clever and quickly adapt, so rules need to be updated.

Remark: Paul Graham of Y Combinator popularized NB classifier (which he calls Bayesian filtering) in [1] as a way to battle spam. It is a fun read. The idea of using NB classifier to battle spam is older though (see [2])

[1]: `http://www.paulgraham.com/spam.html`
[2]: `http://research.microsoft.com/en-us/um/people/horvitz/junkfilter.htm`

NB for text classification

$P(X \mid Y)$ is huge.

- Documents contain many words
- There are many possible words in the vocabulary

The Naïve assumption helps a lot

- $P(X_i = x_i \mid Y = y)$ is simply the probability of observing word x_i in a document on topic y
- $P(\text{"hockey"} \mid Y = \text{sports})$

$$P(y) \cdot \prod_{i=1}^{\text{LengthDoc}} P(x_i \mid y)$$

Bag of words representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet.

x **love** xxxxxxxxxxxxxxxx **sweet** xxxxxxx **satirical** xxxxxxxxxx xxxxxxxxxxx **great** xxxxxxx xxxxxxxxxxxxxxxxxxx **fun** xxxx xxxxxxxxxxxxx **whimsical** xxxx **romantic** xxxx **laughing** xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx xxxxxxxxxxxxxx **recommend** xxxxx xxxxxxxxxxxxxxxxxxxxxxxxxxxxx xx **several** xxxxxxxxxxxxxxxxx xxxxx **happy** xxxxxxxxx **again** xxxxxxxxxxxxxxxxxxxxxxxxxxxx xxxxxxxxxxxxxxxx

Bag of words representation

Position in a document does not matter

$$P(X_i = x_i \mid Y = y) = P(X_k = x_i \mid Y)$$

- "Bag of words" representation ignores the order of words in a document
- Sounds really silly, but often works very well!

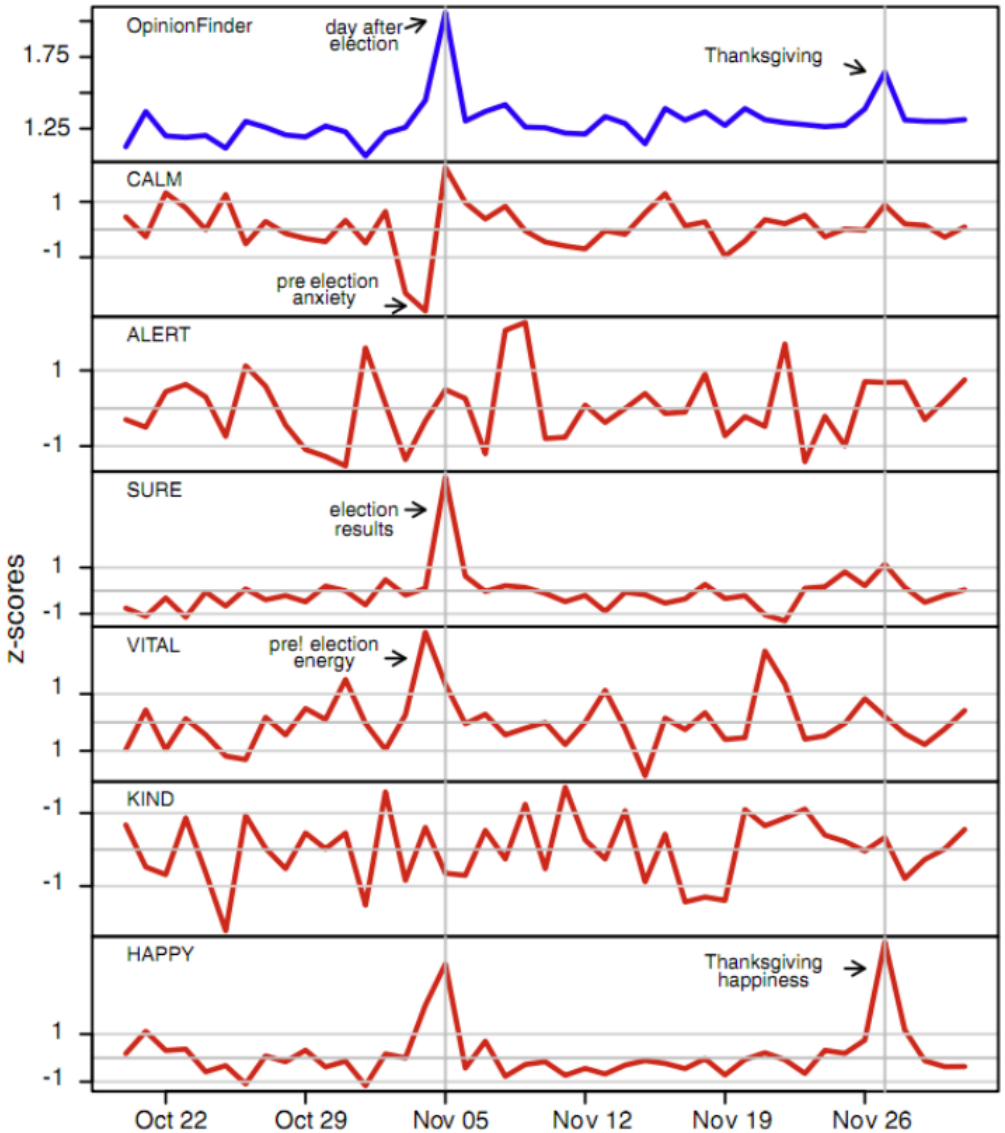
The following two documents are the same

When the lecture is over, remember to wake up the person sitting next to you in the lecture room.

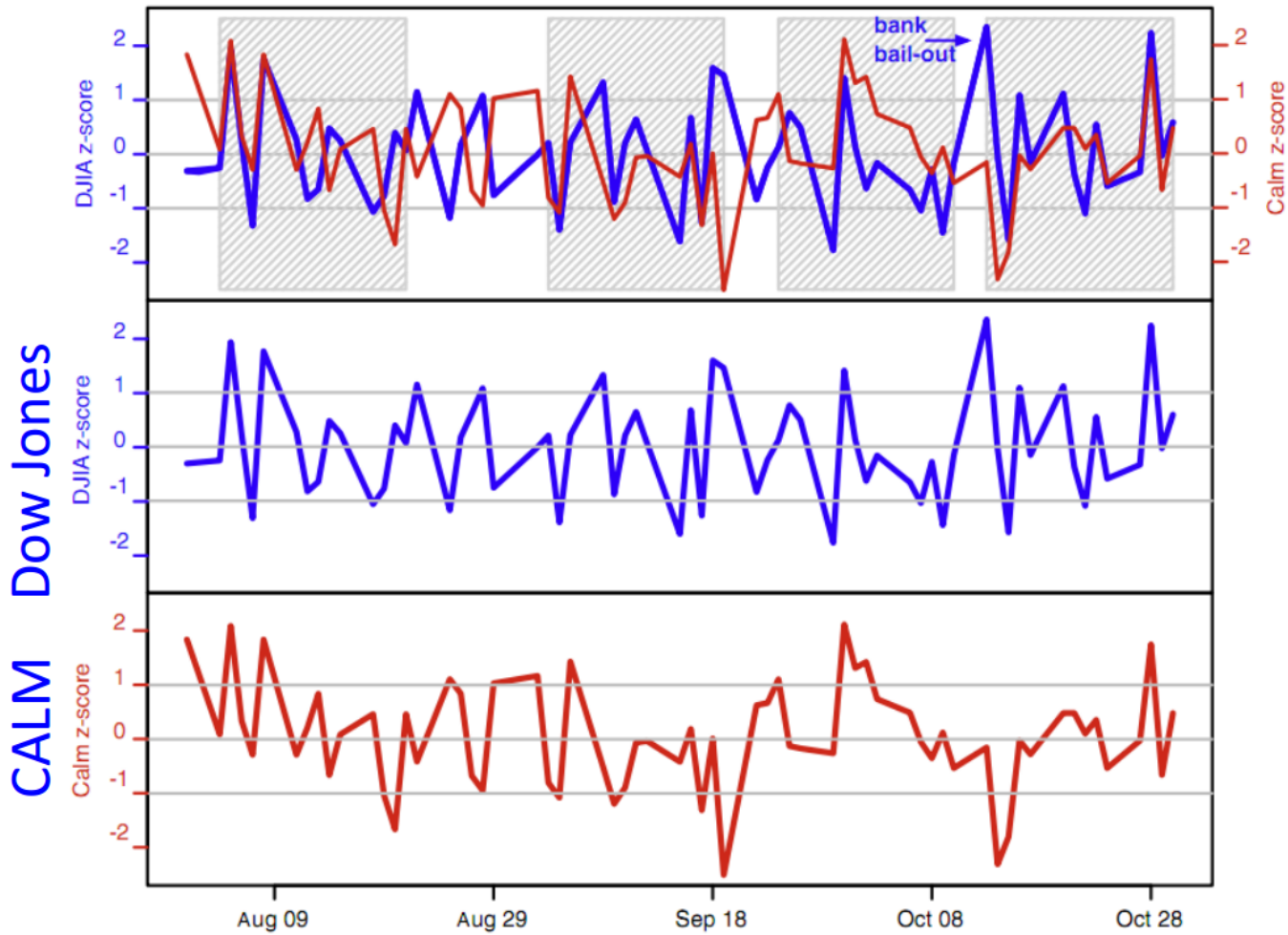
in is lecture lecture next over person remember room sitting the the the to to up wake when you

Sentiment analysis

Twitter mood predicts the stock market. Johan Bollen, Huina Mao, Xiao-Jun Zeng



Sentiment analysis: CALM predicts DJIA 3 days later



Sentiment analysis

R detour: See *04_code.R*

Application of NB to Large Movie Review Dataset.
`http://ai.stanford.edu/~amaas/data/sentiment/index.html`

Practical considerations

We can use pairs of words to better capture meaning of the review.

We can double count words in the subject or title.

- some words carry more information
- domain specific words

Can incorporate additional information, like URLs, senders email, senders domain.

It is common to use all words, except "stop words." These are words that carry no information (for example: a, the, at, which, . . .).

- lists of stop words exist for most languages

Recap: Naïve Bayes classification

Simple classification procedure based on Bayes rule.

Makes a very naïve modeling assumption.

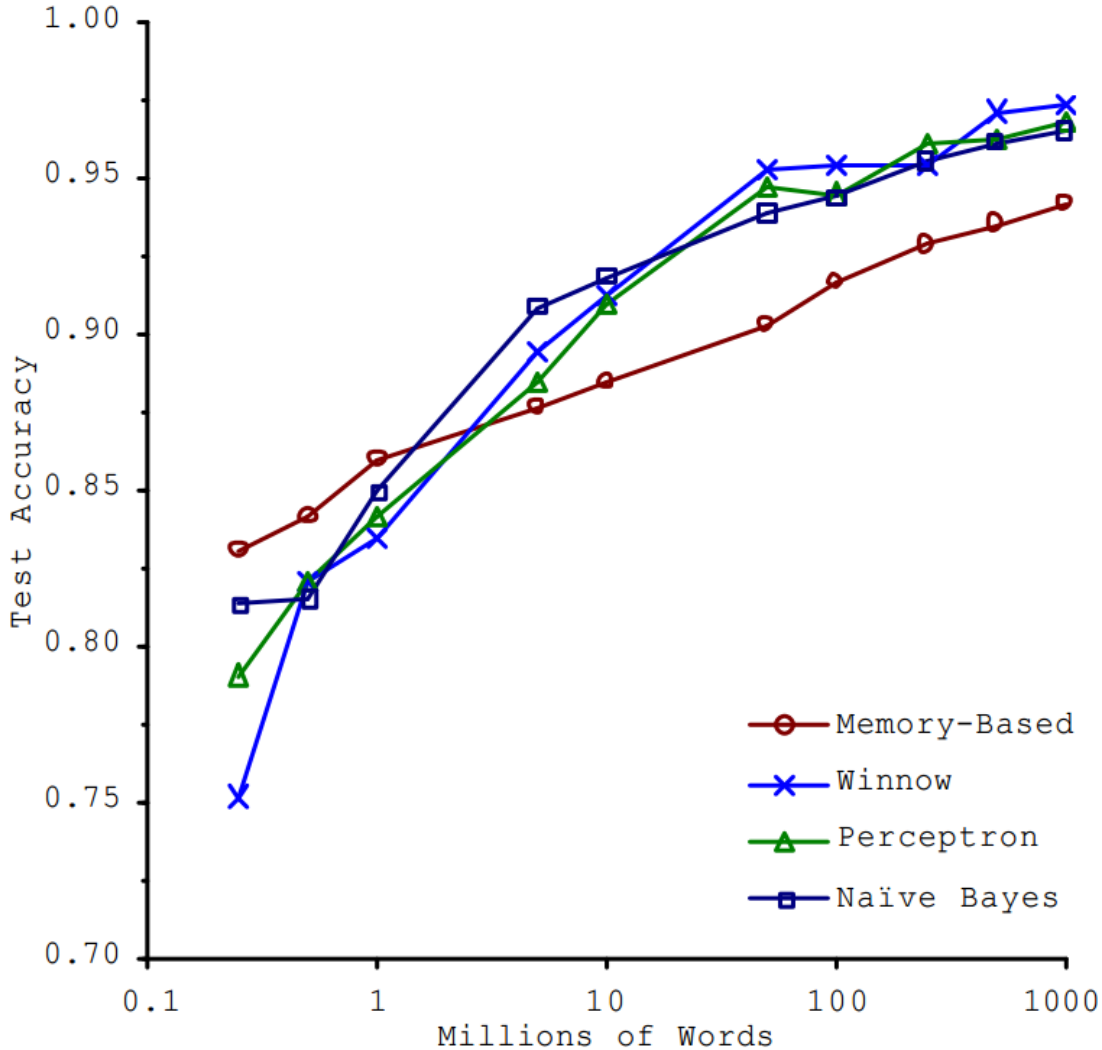
Learning is very simple: counting!

Very fast, low storage requirements.

Simple baseline that can be quickly implemented.

Aside: More data beats fancy algorithms

Confusion set disambiguation problem (e.g., to, two, too)



[Banko and Brill, 2001] http://research.microsoft.com/pubs/66840/acl2001.pdf