Pattern Analysis and Recognition

Lecture 5: probabilities, bayesian inference

Last time on Pattern Analysis and Recognition

RECAP

By now you should know

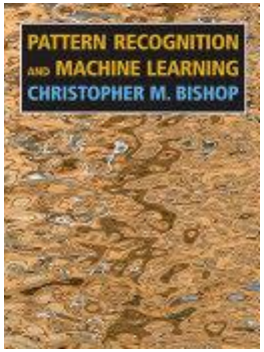
- How to estimate the parameters of a linear or polynomial model using regression and gradient descent
- How to use the core of regression to do classification, by using the logistic (sigmoid) function
- How to normalise your data
- What the variance / bias trade-off is
- How to detect / ensure that you do not overfit your data using regularisation
- How to reduce the dimensionality of your feature space (subspace methods)

PROBABILITIES

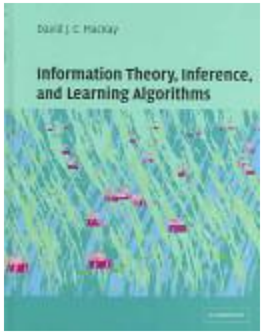
Resources

Some of the material in this slides was borrowed from:

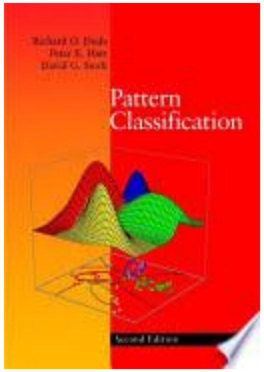
C. Bishop, "Pattern Recognition and Machine Learning", Springer, 2006
Some related material available:
http://research.microsoft.com/en-us/um/people/cmbishop/prml/index.htm



D. MacKay, "Information Theory, Inference and Learning Algorithms", Cambridge University Press, 2003.
Book available online:
http://www.inference.phy.cam.ac.uk/mackay/



R.O. Duda, P.E. Hart, D.G. Stork, "Pattern Classification", Wiley & Sons, 2000
Have a look inside at selected chapters:
http://books.google.es/books/about/Pattern_Classification.html?id=Br33IRC3PkQC&redir_esc=y



Inference

- Inference is the process of deriving conclusions based solely on what one already knows

> all men are mortals
> Socrates is a man
> therefore Socrates is mortal

- **Classic logic** (deterministic reasoning) is concerned with certainty (true / false)
- **Statistical inference** draws conclusions in the presence of uncertainty, generalising deterministic reasoning. Motivations:
  - Statistics: Need to infer parameters or test hypotheses on statistical data in a rigorous, quantitative way
  - AI: Need to reason efficiently about uncertain quantities
- **Bayesian inference** quantifies degrees of belief using probabilities
  - *E.g. P(the ice cap will melt in the next 10 years) = 0.7*
  - Through rules of probability, the probability (degree of certainty) of a conclusion or parameter can be calculated

Searching in the dark

Experiment

Randomly open a drawer and pick a pair of socks.
Replace the item and repeat the procedure many times over

Assumptions

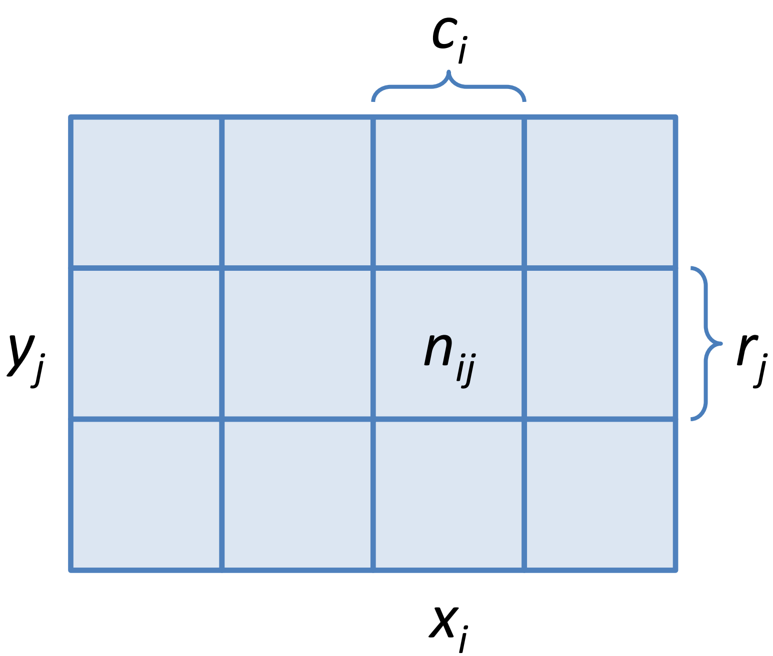
Suppose that we tend to pick from the lower drawer more often, say 60% of the time (40% of the time we pick from the upper drawer)

Definitions

Drawer *D* is a **random variable** that can take one of two possible values *{upper, lower}*. Same for the colour of the socks *C = {black, white}*

Q1. What is the probability of selecting a black pair of socks?
Q2. Given that the socks we picked are white, what is the probability that the drawer we chose from was the upper one?

Probability Theory

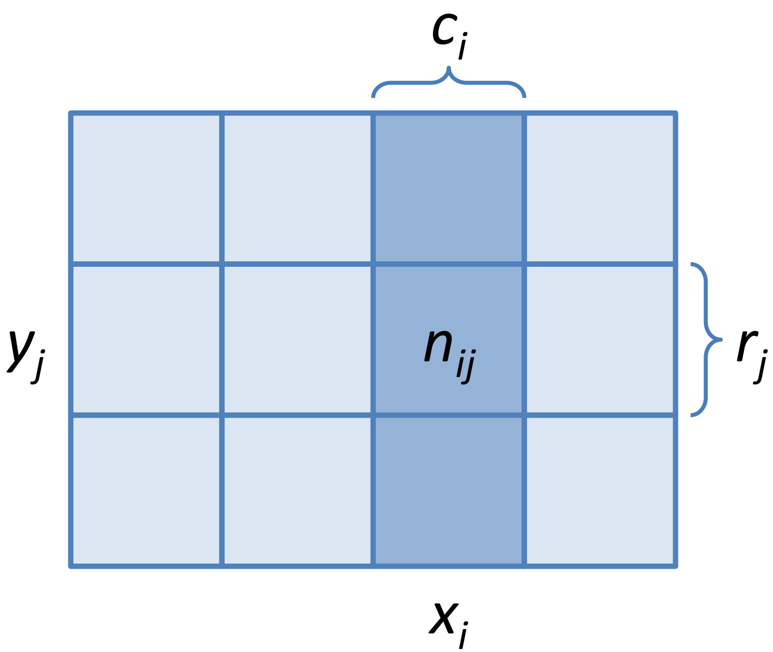


Random variables

$$X = \{x_i\}$$

$$Y = \{y_j\}$$

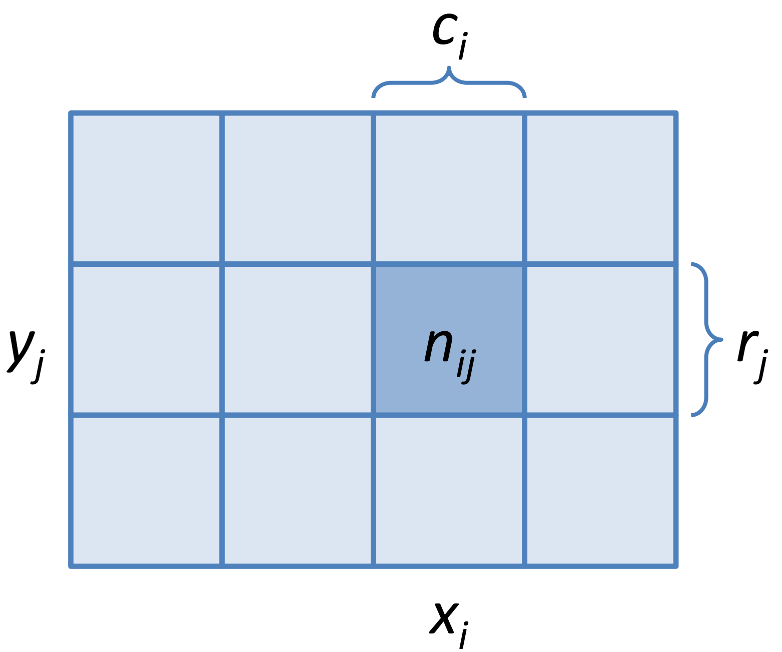
Probability Theory



Marginal Probability

$$p(X = x_i) = \frac{c_i}{N}$$

Probability Theory



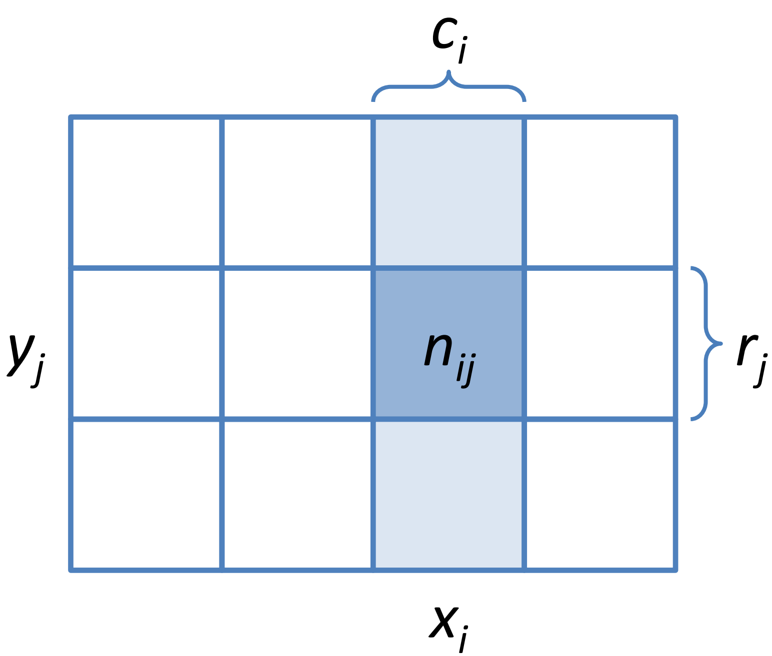
Marginal Probability

$$p(X = x_i) = \frac{c_i}{N}$$

Joint Probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Probability Theory



Marginal Probability

$$p(X = x_i) = \frac{c_i}{N}$$

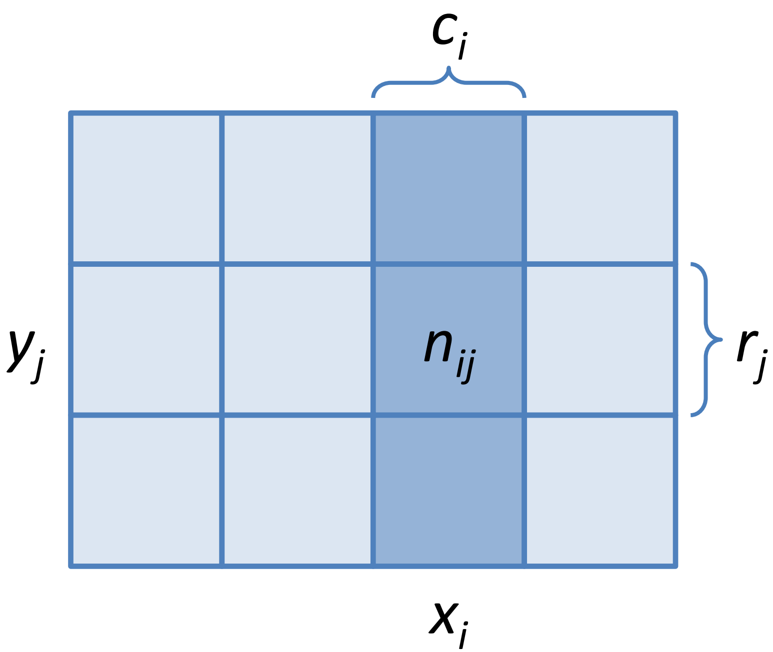
Joint Probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Conditional Probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

Probability Theory

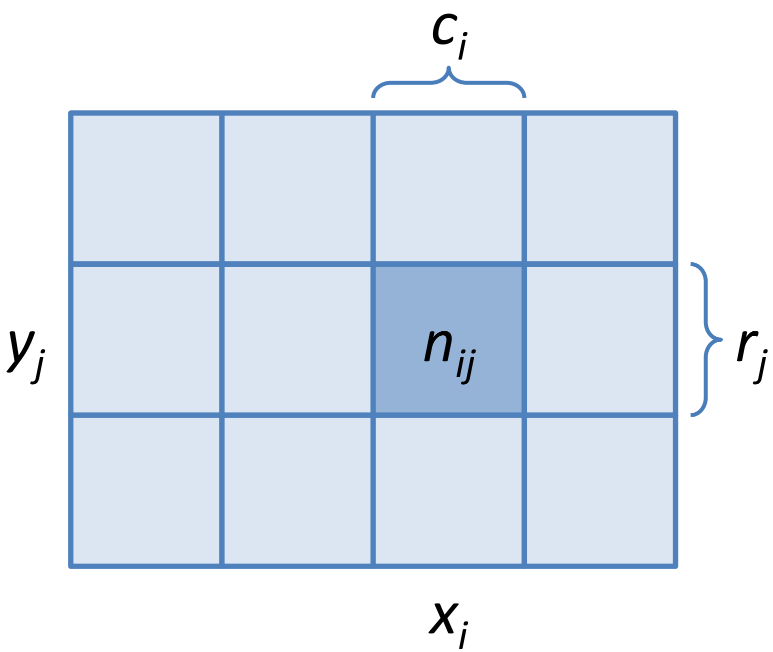


Sum Rule

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N}\sum_{j=1}^{L} n_{ij}$$

$$= \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

Probability Theory



Product Rule

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

$$= \frac{n_{ij}}{c_i} \frac{c_i}{N}$$

$$= p(Y = y_j | X = x_i) p(X = x_i)$$

The Rules of Probability

Sum Rule $$p(X) = \sum_{Y} p(X,Y)$$

Product Rule $$p(X,Y) = p(Y|X)\,p(X)$$

Bayes' Theorem

$$\underbrace{p(Y|X)}_{\text{posterior}} = \frac{\overbrace{p(X|Y)}^{\text{likelihood}}\,\overbrace{p(Y)}^{\text{prior}}}{\underbrace{p(X)}_{\text{evidence}}}$$

$$p(X) = \sum_{Y} p(X|Y)\,p(Y)$$

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Bayes' Theorem

$$p(Y|X) = \frac{p(X|Y)\,p(Y)}{p(X)}$$

$$\mathrm{posterior} = \frac{\mathrm{likelihood} \times \mathrm{prior}}{\mathrm{evidence}}$$

Searching in the dark revisited

Experiment

Randomly open a drawer and pick a pair of socks.
Replace the item and repeat the procedure many times over

Assumptions

Suppose that we tend to pick from the lower drawer more often, say 60% of the time (40% of the time we pick from the upper drawer)

Definitions

Drawer *D* is a **random variable** that can take one of two possible values *{upper, lower}*. Same for the colour of the socks *C = {black, white}*

Q1. What is the probability of selecting a black pair of socks?
Q2. Given that the socks we picked are white, what is the probability that the drawer we chose from was the upper one?

Searching in the dark revisited

Q2. Given that the socks we picked are white, what is the probability that the drawer we chose from was the upper one?

$$p(D=u)=\frac{4}{10} \qquad p(D=l)=\frac{6}{10}$$

$$p(C=b|D=u)=\frac{1}{4} \qquad p(C=b|D=l)=\frac{3}{4}$$

$$p(C=w|D=u)=\frac{3}{4} \qquad p(C=w|D=l)=\frac{1}{4}$$

$$p(C=b)=p(C=b|D=u)\,p(D=u)+p(C=b|D=l)\,p(D=l)$$

$$=\frac{1}{4}\times\frac{4}{10}+\frac{3}{4}\times\frac{6}{10}=\frac{11}{20}$$

$$p(C=w)=1-p(C=b)=1-\frac{11}{20}=\frac{9}{20}$$

$$p(D=u|C=w)=\frac{p(C=w|D=u)\,p(D=u)}{p(C=w)}=\frac{3}{4}\times\frac{4}{10}\times\frac{20}{9}=\frac{2}{3}$$

Likelihood is not a probability

- Likelihood $P(E|\theta)$
  - For fixed θ defines a probability over E
  - For fixed E defines the likelihood of θ

- Do not say “the likelihood of the data” but the “likelihood of the parameters θ”

A Medical Test

Jo has a test for a nasty disease. We denote Jo's state of health by the variable *a* and the test result by *b*.

$a = 0$ Jo does not have the disease	$b = 0$, test is 'negative'
$a = 1$ Jo has the disease	$b = 1$, test is 'positive'

The test is 95% reliable: in 95% of cases of people who really have the disease, a positive result is returned, and in 95% of cases of people who do not have the disease, a negative result is obtained.

1% of people of Jo's age and background have the disease.

Q: Jo has the test, and the result is positive. What is the probability that Jo has the disease?

From MacKay, "Information Theory, Inference, and Learning Algorithms", Cambridge University Press 2003, Chapter 2.2. (book available to download: http://www.inference.eng.cam.ac.uk/mackay/itila/)

A Medical Test

a **is Jo's State**
a = 0, Jo does not have the disease
a = 1, Jo has the disease

b **is the result**
b = 0, test is 'negative'
b = 1, test is 'positive'

$$p(a=1|b=1)=?$$

A Medical Test - Solution

We write down all the provided probabilities. The test reliability specifies the conditional probability of b given a:

$P(b=1 \mid a=1) = 0.95$ $\quad$ $P(b=1 \mid a=0) = 0.05$

$P(b=0 \mid a=1) = 0.05$ $\quad$ $P(b=0 \mid a=0) = 0.95$

and the disease prevalence tells us about the marginal probability of a:

$P(a=1) = 0.01$ $\quad$ $P(a=0) = 0.99$

From the marginal $P(a)$ and the conditional probability $P(b \mid a)$ we can deduce the joint probability $P(a, b) = P(a)P(b \mid a)$ and any other probabilities we are interested in. For example, by the sum rule, the marginal probability of b=1 – the probability of getting a positive result – is

$$P(b=1) = P(b=1 \mid a=1)P(a=1) + P(b=1 \mid a=0)P(a=0)$$

Jo has received a positive result b=1 and is interested in how plausible it is that she has the disease (i.e., that a=1). The man in the street might be duped by the statement 'the test is 95% reliable, so Jo's positive result implies that there is a 95% chance that Jo has the disease', but this is incorrect. The correct solution to an inference problem is found using Bayes' theorem.

$$P(a=1 \mid b=1) = P(b=1 \mid a=1)P(a=1) / [P(b=1 \mid a=1)P(a=1) + P(b=1 \mid a=0)P(a=0)]$$

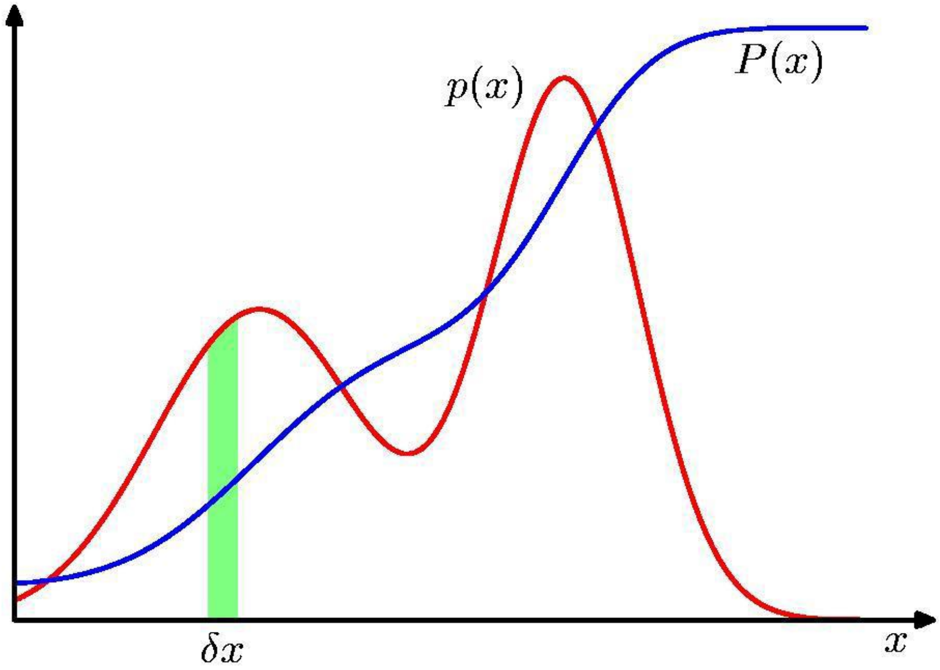
$$= 0.95 \times 0.01 / (0.95 \times 0.01 + 0.05 \times 0.99) = 0.16$$

So in spite of the positive result, the probability that Jo has the disease is only 16%

From MacKay, "Information Theory, Inference, and Learning Algorithms", Cambridge University Press 2003, Chapter 2.2. (book available to download: http://www.inference.eng.cam.ac.uk/mackay/itila/)

Probability Densities

The concept of probability for discrete variables can be extended to that of a probability density over a continuous variable *x*



$$p\big(x \in (a,b)\big) = \int_a^b p(x)dx$$

$$p(x) \geq 0$$

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

Cumulative distribution function: $P(z) = \int_{-\infty}^{z} p(x)dx$

THE MEANING OF PROBABILITY

Frequentist vs Bayesian Viewpoint

Frequentist Viewpoint

- Probabilities describe frequencies or outcomes in random experiments
- Probabilities quantifying beliefs (e.g. priors) are subjective since they depend on assumptions

Bayesian Viewpoint

- Probabilities can be used to describe degrees of belief in propositions that do not involve random variables
- Subjectivity is not a defect: you have to make assumptions before you can do inference
  - “1% of people of Jo’s age and background have the disease
  - The test performs equally well for all people and external conditions
- Making assumptions explicit makes them easier to criticize and modify

Bayesian Viewpoint

- **Uncertainty** is expressed as a probability
  - *This not an ad-hoc choice, it has been shown that if numerical values are used to represent degrees of belief, then a simple set of axioms encoding common sense properties of such beliefs leads uniquely to a set of rules equivalent to the sum and product rules of probability*
- This allows us to see probability theory as an extension of Boolean logic to situations involving uncertainty

Forward and Inverse Probabilities

- Forward probability
  - compute probability distribution or expectation of some quantity that depends on data produced by some known model of a process
  - involves a generative model that describes a process that is assumed to give rise to some data
- Inverse probability
  - instead of computing the probability distribution of some quantity produced by the process compute the conditional probability of **unobserved variables** of the process, given observed variables (data)

Bayesian Inference

- Bayesian inference is an inverse probability problem
- With Bayesian inference we derive conclusions by
  - Assigning probabilities to beliefs
  - Applying Bayes' theorem

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

- In real life we have to make subjective assumptions for instance in order to assign priors or to model the likelihood – Bayesian inference is a principled way whereby all assumptions have to be made explicit

What's Next

- Expectations covariances and the Gaussian Distribution
- Curve fitting revisited and how what we saw in regression follows from the Bayesian approach
- Bayesian classification