Inf2b Learning and Data

Lecture 5: Introduction to statistical pattern recognition

Hiroshi Shimodaira
(Credit: Iain Murray and Steve Renals)

Centre for Speech Technology Research (CSTR)
School of Informatics
University of Edinburgh

Jan-Mar 2014

Today's Schedule

1. Probability (review)
2. What is Bayes' theorem for?
3. Statistical classification

Motivation for probability

In some applications we need to:

- Communicate uncertainty
- Use prior knowledge
- Deal with missing data

(we cannot easily measure similarity)

1. Probability (review)
2. What is Bayes' theorem for?
3. Statistical classification

Warming up

- Throwing two dices
  - Probablity of $\{1, 1\}$?
  - Probablity of $\{2, 5\}$?
- Drawing two cards from a deck of cards
  - Probability of $\{\text{Club}, \text{Spade}\}$?
  - Probability of $\{\text{Club}, \text{Club}\}$?

Warming up

- Probability that a student in Informatics has eyeglasses?
- Probability that you live more than 90 years?
- When a real dice is thrown, is the probability of getting $\{1\}$ $\frac{1}{6}$?

Theoretical probability vs. Empirical probability

Rules of Probability

Sum Rule:

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

Product Rule:

$$\begin{aligned} P(Y{=}y_j, X{=}x_i) &= P(Y{=}y_j \,|\, X{=}x_i)\, P(X{=}x_i) \\ &= P(X{=}x_i \,|\, Y{=}y_j)\, P(Y{=}y_j) \end{aligned}$$

Random variables	Events/values
X	$\{x_1, x_2, \ldots, x_L\}$
Y	$\{y_1, y_2, \ldots, y_L\}$

Rules of Probability

Sum Rule:

$$P(X) = \sum_{Y} P(X, Y)$$

Product Rule:

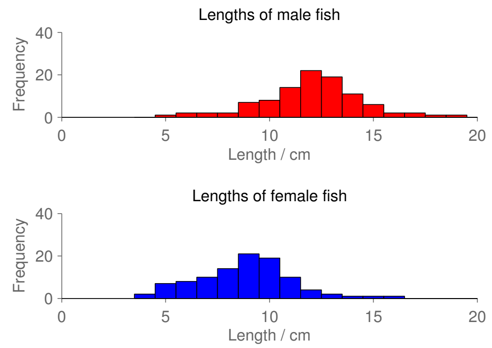
$$\begin{aligned} P(Y, X) &= P(Y \,|\, X)\, P(X) \\ &= P(X \,|\, Y)\, P(Y) \end{aligned}$$

If X and Y are independent,

$$P(X|Y) = P(X)$$
$$P(X, Y) = P(X)P(Y)$$

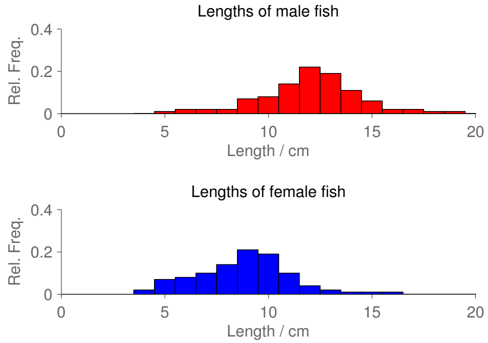
Example: determinig the sex of fish

Histograms of fish lengths



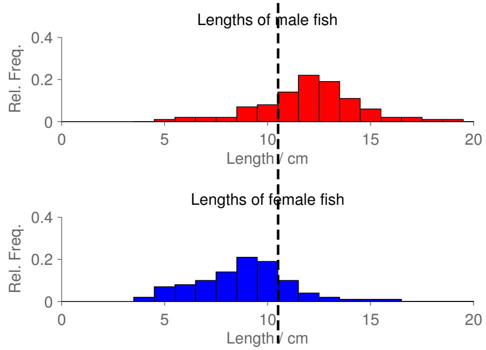
Example: determinig the sex of fish

Relative frequencies of fish length



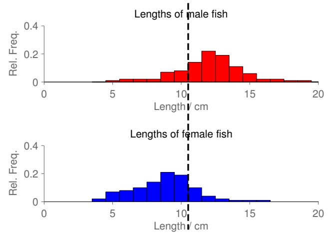
Example: determinig the sex of fish

Possible decision boundary



Fish questions

- How to classify 4 cm, or 19 cm fish?
- How to classify 22 cm fish?
- How to classify 10 cm fish?
- What if there are 10× more male fish than female?
- What if you're forbidden from catching female fish?

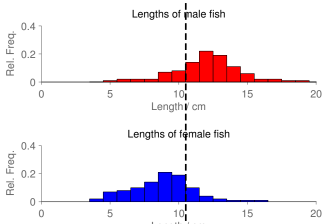


Fish questions

Relative frequeny of male fish length:	$P(X = x \mid C = \mathrm{M})$
Relative frequeny of female fish length:	$P(X = x \mid C = \mathrm{F})$

Given a fish length, $\mathbf{x}$, is it sensible to decide as follows?

male fish if $P(x|\mathrm{M}) > P(x|\mathrm{F})$
female fish if $P(x|\mathrm{M}) < P(x|\mathrm{F})$

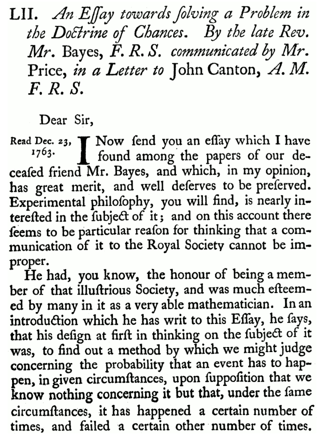


Bayes' Theorem

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

Tohmas Bayes (?) (1701? – 1761)

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



LII. *An Eſſay towards ſolving a Problem in the Doctrine of Chances. By the late Rev. Mr.* Bayes, *F. R. S. communicated by Mr.* Price, *in a Letter to* John Canton, *A. M. F. R. S.*

Dear Sir,

Read Dec. 23, 1763. I Now ſend you an eſſay which I have found among the papers of our deceaſed friend Mr. Bayes, and which, in my opinion, has great merit, and well deſerves to be preſerved. Experimental philoſophy, you will find, is nearly intereſted in the ſubject of it; and on this account there ſeems to be particular reaſon for thinking that a communication of it to the Royal Society cannot be improper.

He had, you know, the honour of being a member of that illuſtrious Society, and was much eſteemed by many in it as a very able mathematician. In an introduction which he has writ to this Eſſay, he ſays, that his deſign at firſt in thinking on the ſubject of it was, to find out a method by which we might judge concerning the probability that an event has to happen, in given circumſtances, upon ſuppoſition that we know nothing concerning it but that, under the ſame circumſtances, it has happened a certain number of times, and failed a certain other number of times.

'Bayesian' philosophy refs

Non-examinable!

Bayes' paper:
http://www.jstor.org/stable/105741
http://dx.doi.org/10.1093/biomet/45.3-4.296 (re-typeset)

Cox's paper:
http://dx.doi.org/10.1119/1.1990764
http://dx.doi.org/10.1016/S0888-613X(03)00051-3 modern commentary

MacKay textbook, amongst many others

Card prediction

3 cards with coloured faces:

1. one white and one black face
2. two black faces
3. two white faces

I shuffle cards and turn them over randomly. I select a card and way-up uniformly at random and place it on a table.

Question: You see a white face. What is the probability that the other side of the same card is black? You see a black face. What is the probability that the other side of the same card is white?

A) $1/4$ **B)** $1/3$ **C)** $1/2$ **D)** $2/3$ **E)** $3/4$ **Z)** ?

Notes on the card prediction problem:

This card problem is Ex. 8.10a), MacKay, p142.
It is *not* the same as the famous 'Monty Hall' puzzle: Ex. 3.8–9 and
http://en.wikipedia.org/wiki/Monty_Hall_problem
The Monty Hall problem is also worth understanding. Although the card problem is (hopefully) less controversial and more straightforward. The process by which a card is selected should be clear: $P(c) = 1/3$ for $c = 1, 2, 3$, and the face you see first is chosen at random: e.g., $P(x_1 = \mathrm{B} \mid c = 1) = 0.5$.
Many people get this puzzle wrong on first viewing (it's easy to mess up). If you do get the answer right immediately (are you sure?), this is will be a simple example on which to demonstrate some formalism.

How do we solve it formally?

Use Bayes theorem?

$$P(x_2 = \mathtt{W} \mid x_1 = \mathtt{B}) = \frac{\boxed{P(x_1 = \mathtt{B} \mid x_2 = \mathtt{W})}\, P(x_2 = \mathtt{W})}{P(x_1 = \mathtt{B})}$$

The $\boxed{\text{boxed}}$ term is no more obvious than the answer!

Bayes theorem is used to 'invert' forward generative processes that we understand.
The first step to solve inference problems is to write down a model of your data.

The card game model

Cards: 1) $\mathtt{B}|\mathtt{W}$, 2) $\mathtt{B}|\mathtt{B}$, 3) $\mathtt{W}|\mathtt{W}$

$$P(c) = \begin{cases} 1/3 & c = 1, 2, 3 \\ 0 & \text{otherwise.} \end{cases}$$

$$P(x_1 = \mathtt{B} \mid c) = \begin{cases} 1/2 & c = 1 \\ 1 & c = 2 \\ 0 & c = 3 \end{cases}$$

Bayes theorem can 'invert' this to tell us $P(c \mid x_1 = \mathtt{B})$;
infer the generative process for the data we have.

Inferring the card

Cards: 1) $\mathtt{B}|\mathtt{W}$, 2) $\mathtt{B}|\mathtt{B}$, 3) $\mathtt{W}|\mathtt{W}$

$$\begin{aligned} P(c \mid x_1 = \mathtt{B}) &= \frac{P(x_1 = \mathtt{B} \mid c)\, P(c)}{P(x_1 = \mathtt{B})} \\ &\propto \begin{cases} 1/2 \cdot 1/3 = 1/6 & c = 1 \\ 1 \cdot 1/3 = 1/3 & c = 2 \\ 0 & c = 3 \end{cases} \\ &= \begin{cases} 1/3 & c = 1 \\ 2/3 & c = 2 \end{cases} \end{aligned}$$

Q: *"But aren't there two options given a black face, so it's 50–50?"*

A: There are two options, but the likelihood for one of them is $2\times$ bigger

Predicting the next outcome

(This slide and the next are not really required for this course.)

For this problem we can spot the answer, for more complex problems we want a formal means to proceed.
$P(x_2 \mid x_1 = \mathtt{B})$?
Need to introduce c to use expressions we know:

$$\begin{aligned} P(x_2 \mid x_1 = \mathtt{B}) &= \sum_{c \in 1,2,3} P(x_2, c \mid x_1 = \mathtt{B}) \\ &= \sum_{c \in 1,2,3} P(x_2 \mid x_1 = \mathtt{B}, c)\, P(c \mid x_1 = \mathtt{B}) \end{aligned}$$

Predictions we would make if we knew the card, weighted by the posterior probability of that card. $P(x_2 = \mathtt{W} \mid x_1 = \mathtt{B}) = 1/3$

Strategy for solving any inference and prediction problem:

When interested in something y, we often find we can't immediately write down mathematical expressions for $P(y \mid \text{data})$.
So we introduce stuff, z, that helps us define the problem:

$$P(y \mid \text{data}) = \sum_z P(y, z \mid \text{data})$$

by using the sum rule. And then split it up:

$$P(y \mid \text{data}) = \sum_z P(y \mid z, \text{data})\, P(z \mid \text{data})$$

using the product rule. If knowing extra stuff z we can predict y, we are set: weight all such predictions by the posterior probability of the stuff ($P(z \mid \text{data})$, found with Bayes theorem).
Sometimes the extra stuff summarizes everything we need to know to make a prediction:

$$P(y \mid z, \text{data}) = P(y \mid z)$$

although not in the formulation of the card game above.

Not convinced?

Not everyone believes the answer to the card game question.
Sometimes probabilities are counter-intuitive. I'd encourage you to write simulations of these games if you are at all uncertain. Here is an Octave/Matlab simulator I wrote for the card game question:

```
cards = [1 1;
         0 0;
         1 0];
num_cards = size(cards, 1);

N = 0; % Number of times first face is black
kk = 0; % Out of those, how many times the other side is white

for trial = 1:1e6
    card = ceil(num_cards * rand());
    face = 1 + (rand < 0.5);
    other_face = (face==1) + 1;
    x1 = cards(card, face);
    x2 = cards(card, other_face);

    if x1 == 0
        N = N + 1;
        kk = kk + (x2 == 1);
    end
end

approx_probability = kk / N
```


1 Probability (review)

2 What is Bayes' theorem for?

3 Statistical classification

Bayes and pattern recognition

Class $C = c_1, \ldots, c_K$; input features $X = \mathbf{x}$

Most probable class: (maximum posterior class)

$$c^* = \arg\max_{c_k} P(c_k \mid \mathbf{x})$$

where

$$\overbrace{P(c_k \mid \mathbf{x})}^{\text{posterior}} = \frac{\overbrace{P(\mathbf{x} \mid c_k)}^{\text{likelihood}}\, \overbrace{P(c_k)}^{\text{prior}}}{P(\mathbf{x})}$$

Might also **minimize expected loss:** (non-examinable)

$$c^* = \arg\min_{c_k} \sum_{c_t} P(c_t \mid \mathbf{x})\, L(c_k, c_t)$$

Posterior probability

Can compute denominator with sum rule:

$$P(\mathbf{x}) = \sum_\ell P(\mathbf{x} \mid c_\ell)\, P(c_\ell)$$

However $P(\mathbf{x})$ is the same for all classes:

$$P(c_k \mid \mathbf{x}) \propto P(\mathbf{x} \mid c_k)\, P(c_k)$$

Choosing between two classes, only requires the *odds*, the ratio of the posterior probabilities.

Back to the fish

100 example male fish, 100 examples females
Estimating likelihoods as relative frequencies gives:

x	$P(x \mid M)$	$P(x \mid F)$	x	$P(x \mid M)$	$P(x \mid F)$
4	0.00	0.02	13	0.19	0.02
5	0.01	0.07	14	0.11	0.01
6	0.02	0.08	15	0.06	0.01
7	0.02	0.10	16	0.02	0.01
8	0.02	0.14	17	0.02	0.00
9	0.07	0.21	18	0.01	0.00
10	0.08	0.19	19	0.01	0.00
11	0.14	0.10	20	0.00	0.00
12	0.22	0.04			

Marginal probabilities

$$P(x) = \sum_{C \in \{M,F\}} P(x \mid C)\, P(C)$$

Requires prior probabilities: $P(C=M)$, $P(C=F)$

If $P(C=M) = P(C=F) = 0.5$: $\quad P(x=6) = 0.05$
(Males and females equally common)

If $P(C=M) = 0.8$: $\quad P(x=6) = 0.024$
($P(C=F)$ must be 0.2; $4\times$ more males than females)

Inferring labels for $x=11$

Equal prior probabilities: classify it as male:

$$\frac{P(M \mid x)}{P(F \mid x)} = \frac{P(x \mid M)\, P(M)}{P(x \mid F)\, P(F)} = \frac{0.14 \cdot 0.5}{0.10 \cdot 0.5} = 1.4$$

Twice as many females as males: (i.e., $P(M) = 1/3$, $P(F) = 2/3$)

$$\frac{P(M \mid x)}{P(F \mid x)} = \frac{P(x \mid M)\, P(M)}{P(x \mid F)\, P(F)} = \frac{0.14 \cdot 1/3}{0.10 \cdot 2/3} = 0.7$$

Classify it as female
M:F is 0.7:1, that is, $P(M \mid x) = 0.7/(0.7+1) \approx 0.41$

Some more questions

Assume $P(M) = P(F) = 0.5$

1. What is the value of $P(M \mid X=4)$?
2. What is the value of $P(F \mid X=18)$?
3. You observe data point $x=20$.
   To which class should it be assigned?

Work through the notes!

Remember to review the notes...
...work through the fruit box example

Similar material in this online text:
http://www.greenteapress.com/thinkbayes/html/thinkbayes002.html