Inf2b - Learning

Lecture 5: Introduction to statistical pattern recognition and Optimisation

Hiroshi Shimodaira
(Credit: Iain Murray and Steve Renals)

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

http://www.inf.ed.ac.uk/teaching/courses/inf2b/
https://piazza.com/ed.ac.uk/spring2020/infr08028
Office hours: Wednesdays at 14:00-15:00 in IF-3.04

Jan-Mar 2020

Today's Schedule

1. Probability (review)
2. What is Bayes' theorem for?
3. Bayes decision rule
4. More about probability
5. Optimisation problems

Motivation for probability

In some applications we need to:

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

(we cannot easily measure similarity)

Warming up

- Throwing two dices
  - Probability of $\{1, 1\}$?
  - Probability 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 (*cont.*)

- 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

aka:

relative frequency

experimental probability

for a sample set drawn from a population

Rules of Probability

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

Product Rule:

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

Abbreviation:

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

X and Y are *independent* iff:

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

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

Rules of Probability (*cont.*)

Sum Rule:

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

Abbreviation:

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

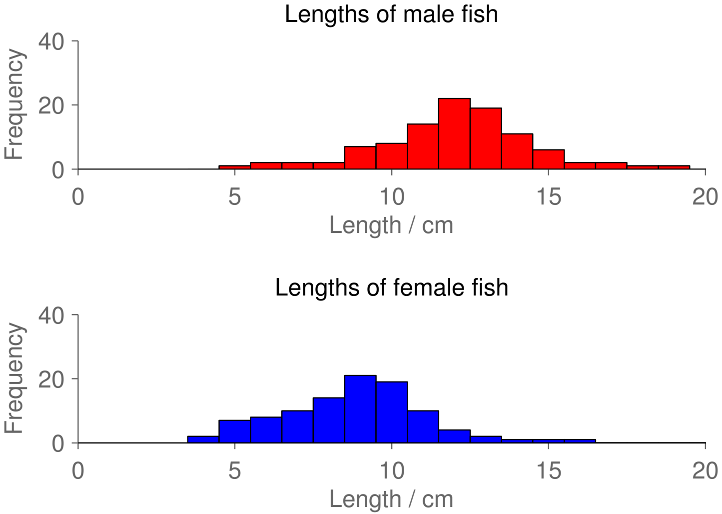
RHS: *Mariginalisation* of the joint probability over Y.
LHS: *Marginal probability* of X.

Application:

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

Example: determining the sex of fish

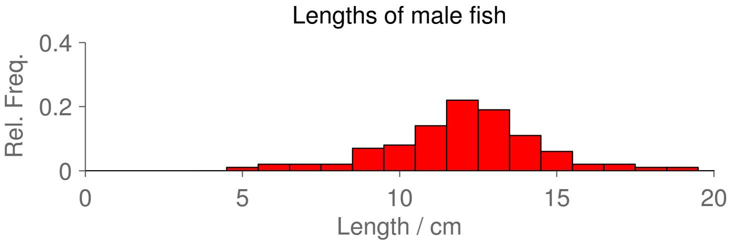
Histograms of fish lengths ($N_F = N_M = 100$)

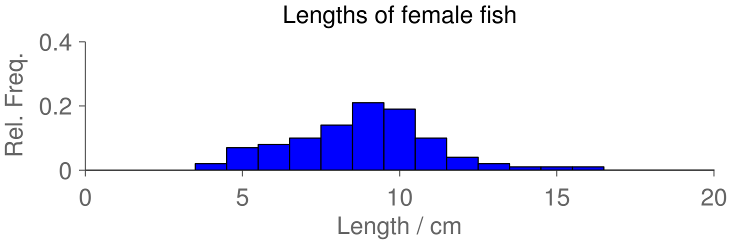


(NB: different example from the one in Note 5.)

Example: determining the sex of fish

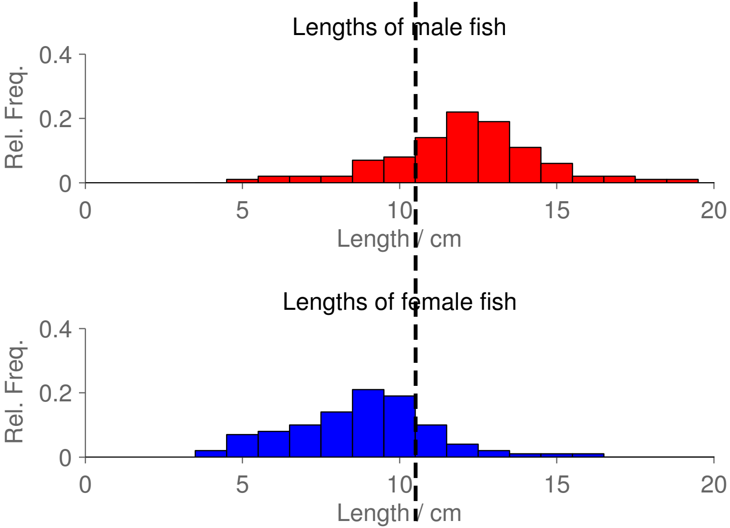
Relative frequencies of fish length





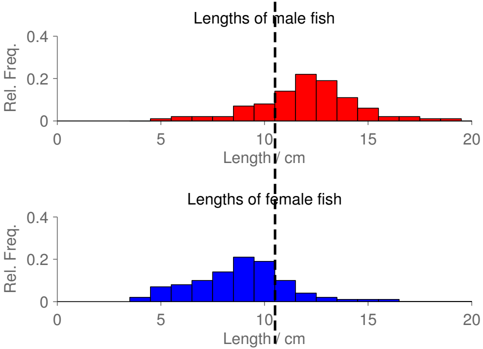
Example: determining the sex of fish

Possible decision boundary



Fish questions

- How to classify 4 cm, or 19 cm fish?
- How to classify 10 cm fish?

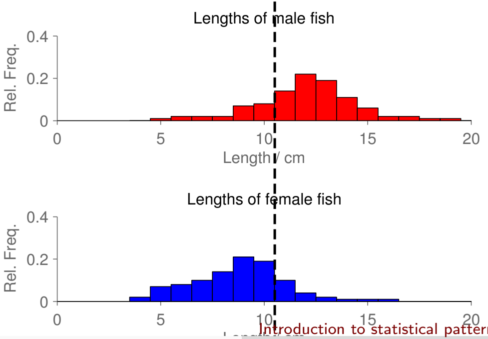


Fish questions

Relative frequency of male fish length:	$P(x \mid \mathrm{M})$
Relative frequency of female fish length:	$P(x \mid \mathrm{F})$

Given a fish length, **x**, is it sensible to decide as follows?

$$\text{If } P(x \mid \mathrm{M}) > P(x \mid \mathrm{F}) \quad \Rightarrow \text{ male fish}$$
$$\text{If } P(x \mid \mathrm{M}) < P(x \mid \mathrm{F}) \quad \Rightarrow \text{ female fish}$$



Fish questions (*cont.*)

How to obtain $P(Y \mid x)$? (where $Y = \{F, M\}$)

- The product rule:

$$\begin{aligned} P(Y, x) &= P(Y \mid x)\, P(x) \\ &= P(x \mid Y)\, P(Y) \end{aligned}$$

- Posterior probabilities:

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

i.e.

$$\begin{aligned} P(M \mid x) &= \frac{P(x \mid M)\, P(M)}{P(x)} \propto P(x \mid M)\, P(M) \\ P(F \mid x) &= \frac{P(x \mid F)\, P(F)}{P(x)} \propto P(x \mid F)\, P(F) \end{aligned}$$

Bayes' Theorem

$$P(H \mid E) = \frac{P(E \mid H)\, P(H)}{P(E)}$$

Thomas Bayes (?) (1702? – 1761)

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

c.f. Bayesian inference, Bayesian

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

Bayes decision rule

Class $C = \{1, \ldots, K\}$; C_k to denote $C = k$; input features $X = \mathbf{x}$

Choose the most probable class: (maximum posterior class)

$$k_{\max} = \arg\max_{k \in C} P(C_k \mid \mathbf{x}) = \arg\max_{k} P(\mathbf{x} \mid C_k)\, P(C_k)$$

where

$$\overbrace{P(C_k \mid \mathbf{x})}^{\text{posterior}} = \frac{\overbrace{P(\mathbf{x} \mid C_k)}^{\text{likelihood}}\ \overbrace{P(C_k)}^{\text{prior}}}{P(\mathbf{x})} = \frac{P(\mathbf{x} \mid C_k)\, P(C_k)}{\sum_{j=1}^{K} P(\mathbf{x} \mid C_j)\, P(C_j)}$$

$\Rightarrow$

- It is known this decision rule gives minimum error rate. (We will discuss this in Lecture 10)
- Also called
  - Minimum error (misclassification) rate classification (PRML C. M. Bishop (2006) Section 1.5)
  - Maximum posterior probability (MAP) decision rule

Inferring labels for $x = 11$

- **Equal prior probabilities:**

$$
\begin{aligned}
P(M \mid x = 11) &= \frac{P(x = 11 \mid M)\, P(M)}{P(x = 11)} \\
&= \frac{P(x = 11 \mid M) P(M)}{P(x = 11 \mid M)\, P(M) + P(x = 11 \mid F) P(F)} \\
&= \frac{0.14 \cdot 0.5}{0.14 \cdot 0.5 + 0.10 \cdot 0.5} = \frac{0.14}{0.24} = 0.58\dot{3} \\
P(F \mid x = 11) &= \frac{P(x = 11 \mid F)\, P(F)}{P(x = 11 \mid M)\, P(M) + P(x = 11 \mid F)\, P(F)} \\
&= \frac{0.10 \cdot 0.5}{0.14 \cdot 0.5 + 0.10 \cdot 0.5} = \frac{0.10}{0.24} = 0.41\dot{6}
\end{aligned}
$$

$\rightarrow$ classify it as male

NB: For classification, no need to calculate $P(x = 11)$.

Inferring labels for $x = 11$ (*cont.*)

- **Equal prior probabilities:**

$$\frac{P(M \mid x = 11)}{P(F \mid x = 11)} = \frac{P(x = 11 \mid M)\, P(M)}{P(x = 11 \mid F)\, P(F)} = \frac{0.14 \cdot 0.5}{0.10 \cdot 0.5} = 1.4$$

Classify it as male:

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

$$\frac{P(M \mid x = 11)}{P(F \mid x = 11)} = \frac{P(x = 11 \mid M)\, P(M)}{P(x = 11 \mid F)\, P(F)} = \frac{0.14 \cdot 1/3}{0.10 \cdot 2/3} = 0.7$$

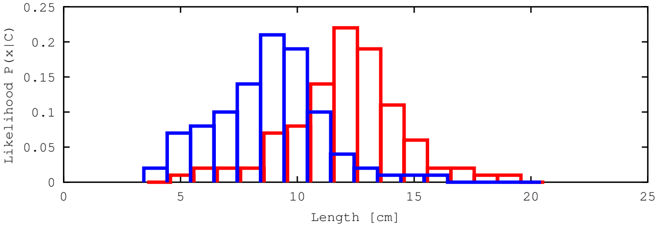
Classify it as female

Likelihood vs posterior probability

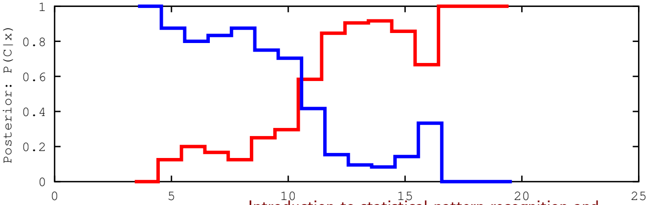
$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)} = \frac{P(\mathbf{x} \mid C_k)\, P(C_k)}{\sum_{j=1}^{K} P(\mathbf{x} \mid C_j)\, P(C_j)}$$

$$P(M) : P(F) = 1 : 1$$

$P(x|C_k)$



$P(C_k|x)$

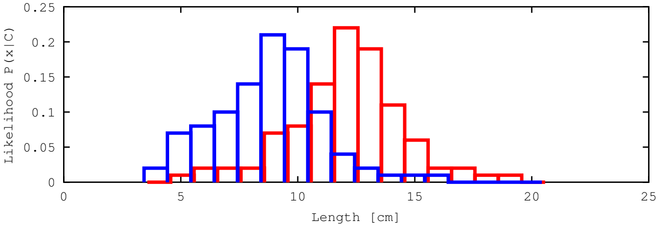


Likelihood vs posterior probability (*cont.*)

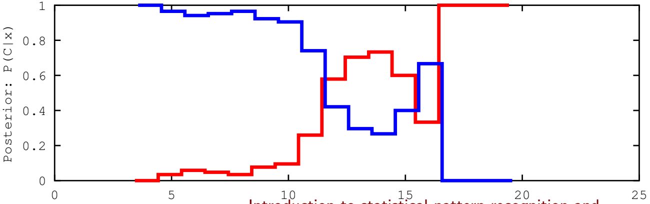
$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)} = \frac{P(\mathbf{x} \mid C_k)\, P(C_k)}{\sum_{j=1}^{K} P(\mathbf{x} \mid C_j)\, P(C_j)}$$

$$P(M) : P(F) = 1 : 4$$

$P(x|C_k)$



$P(C_k|x)$



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 $x = 22$.
     To which class should it be assigned?
- Discuss how you could improve classification performance.
  - What if we increase the number of histogram bins?
  - What if we increase the number of samples?
  - What if we measure not only fish length but also weight?
    (How can we estimate probabilities?)
- It seems that we can estimate $P(C|x)$ directly from data, right?

More about probability

- Conditional probability of three variables

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

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

- Chain rule

$$\begin{aligned} P(X_1, X_2, \ldots, X_N) = \; & P(X_1) P(X_2|X_1) P(X_3|X_1, X_2) \cdots \\ & \cdots P(X_N|X_1, \ldots, X_{N-1}) \end{aligned}$$

Prove!

Independence vs zero correlation

- Independence vs Pearson correlation coefficient $\rho = 0$

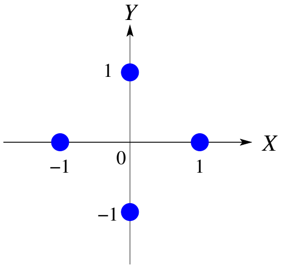
  If X and Y are independent, $\rho_{XY} = 0$.

  The converse is not true.

  See https://en.wikipedia.org/wiki/Correlation_and_dependence

E.g. $(X, Y) = (-1, 0), (0, -1), (0, 1), (1, 0)$, each of which occurs with a probability of $\frac{1}{4}$.

$$\begin{aligned} P(X=-1)\,P(Y=0) &= 1/4 \cdot 1/2 = 1/8 \\ P(X=0)\,P(Y=-1) &= 1/2 \cdot 1/4 = 1/8 \\ P(X=0)\,P(Y=1) &= 1/2 \cdot 1/4 = 1/8 \\ P(X=1)\,P(Y=0) &= 1/4 \cdot 1/2 = 1/8 \end{aligned}$$



$\rho_{XY} = 0$, but $P(X, Y) \neq P(X)\,P(Y)$

i.e., not independent

Optimisation problems we've seen

- Bayes decision rule (MAP decision rule)
$$k_{\max} = \arg\max_{k \in C} P(C_k \,|\, \mathbf{x})$$
- K-NN classification
$$c(\mathbf{z}) = \arg\max_{j \in \{1,\ldots,C\}} \sum_{(\mathbf{x},c) \in U_k(\mathbf{z})} \delta_{j\,c}$$
where $U_k(\mathbf{z})$ is the set of k nearest training examples to $\mathbf{z}$.
- K-means clustering
$$\min_{\{\mathbf{m}_k\}_1^K} E$$
where $E = \frac{1}{N}\sum_{k=1}^{K}\sum_{n=1}^{N} z_{kn}\|\mathbf{x}_n - \mathbf{m}_k\|^2$
- Dimensionality reduction to 2D with PCA
$$\max_{\mathbf{u},\mathbf{v}} \operatorname{Var}(y) + \operatorname{Var}(z)$$
subject to $\|\mathbf{u}\| = 1,\ \|\mathbf{v}\| = 1,\ \mathbf{u} \perp \mathbf{v}$

Optimisation problems : other examples

- Find the shortest path between Edinburgh and London
- Find the cheapest flights from Edinburgh to Tokyo
- For UG4 projects, find the optimal allocation of supervisors and students under given constraints (e.g. no supervisors can take more than five students.)

Types of optimisation problems

- Continuous vs Discrete optimisation
- Unconstrained vs Constrained optimisation

https://neos-guide.org/optimization-tree
https://en.wikipedia.org/wiki/Optimization_problem

Minimisation of *objective function*

$$\min_{\mathbf{x}} f(\mathbf{x}) \qquad \text{where } \mathbf{x} \in \mathcal{R}^D,\ f : \mathcal{R}^D \to \mathcal{R}$$

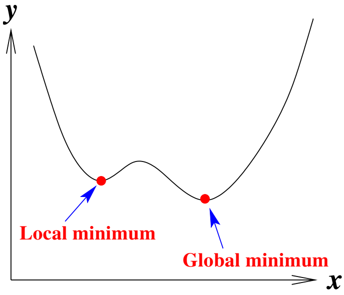
Optimal solution, $\mathbf{x}^* : f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{R}^D$, satisfies †

$$\frac{\partial f(\mathbf{x})}{\partial x_i} = 0,\ \text{for } i = 1, \ldots, D$$

Vector rerpresentation:

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_D} \right)^T = \mathbf{0}$$

$$\text{where } \mathbf{0} = (0, \ldots, 0)^T$$



† This is not a sufficient condition, but a necessary condition.

Optimisation of a quadratic function of one variable

Optimisation problem:

$$\min_x f(x)$$

$$f(x) = ax^2 + bx + c, \quad a > 0$$

- Approach 1:

$$ax^2 + bx + c = a\left(x + \frac{b}{2a}\right)^2 - \frac{b^2}{4a} + c$$

- Approach 2:

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x} = 2ax + b = 0$$

- Solution: $x = -\dfrac{b}{2a}$

Optimisation of a quadratic function of two variables

- Optimisation problem:

$$\min_{\{x,y\}} g(x,y)$$

$$g(x,y) = ax^2 + by^2 + cxy + dx + ey + f$$

$$\text{where } a > 0,\ b > 0,\ c^2 < 4ab$$

$$\rightarrow \quad \frac{\partial g}{\partial x} = 2ax + cy + d = 0$$

$$\frac{\partial g}{\partial y} = 2by + cx + e = 0$$

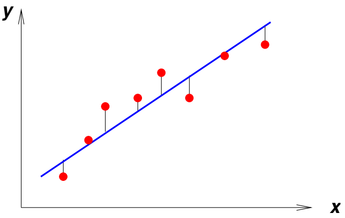
$$\begin{pmatrix} 2a & c \\ c & 2b \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -d \\ -e \end{pmatrix}$$

Least square error line fitting

- Optimisation problem

$$\min_{a,b} \frac{1}{N} \sum_{n=1}^{N} (\hat{y}_n - y_n)^2$$

$$\hat{y}_n = a x_n + b$$



$\rightarrow$

$$\frac{\partial E}{\partial a} = \frac{2}{N} \sum_{n=1}^{N} (a x_n + b - y_n) x_n = 0$$

$$\frac{\partial E}{\partial b} = \frac{2}{N} \sum_{n=1}^{N} (a x_n + b - y_n) = 0$$

$\Rightarrow$ See the lecture note for details.

Least square error line fitting *(cont.)*

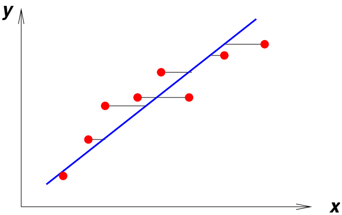
Exercise:

- Optimisation problem

$$\min_{c,d} \frac{1}{N} \sum_{n=1}^{N} (\hat{x}_n - x_n)^2$$

$$\hat{x}_n = c\, y_n + d$$

Find the solution



Iterative optimisation

Many optimisation problems do not have a closed-form solution! (e.g. K-means clustering)

Iterative optimisation method

- Step 1: Choose an initial point $\mathbf{x}_0$, and make $t = 0$.
- Step 2: Choose $\mathbf{x}_{t+1}$ based on an update formula for $\mathbf{x}_t$.
- Step 3: $t \leftarrow t + 1$ and go to step 2 unless stopping criterion is met.

Example of iterative optimisation methods

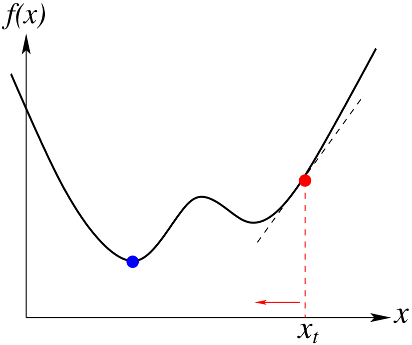
- Gradient descent

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x})\big|_{\mathbf{x}=\mathbf{x}_t} \qquad \text{where } \eta > 0$$

- Conjugate gradient method
- Newton's method

Gradient descent

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \, \nabla f(\mathbf{x})\big|_{\mathbf{x}=\mathbf{x}_t} \qquad \text{where } \eta > 0$$



Things to consider:

- Choice of η (i.e. learning parameter)
- Local-minimum problem

Summary

- *Bayes' theorem* for statistical pattern classification
- Posterior is proportional to prior times likelihood
- $P(x)$ can be obtained with *marginalisation* of $P(x|C)P(C)$
- *Bayes decision rule* achieves minimum error rate classification
- Discuss possible difficulties of applying the Bayes' decision rule to real problems
- Pattern recognition as optimisation problem
- Most of optimisation problem does not have a closed-form solution $\rightarrow$ Iterative optimisation method
- Check the examples in slides, and try the exercises in Note 5.

Mid-course feedback

Your Learn course webpage
$\rightarrow$ (on the left black tab) Have Your Say
$\rightarrow$ Mid-course feedback