Today’s Schedule

- Probability (review)
- What is Bayes’ theorem for?
- Bayes decision rule
- More about probability
- 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 {Club, Spade}? 
  - Probability of {Club, 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} ?

Example: determining the sex of fish

Random variables

<table>
<thead>
<tr>
<th>Events/values</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>{x1, x2, ... , xL}</td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>{y1, y2, ... , yM}</td>
<td></td>
</tr>
</tbody>
</table>

Product Rule:

\[ P(Y = y_j | X = x_i) = P(Y = y_j) P(X = x_i) / P(Y) \]

Abbreviation:

\[ P(Y | X) = P(Y | X) P(X) / P(Y) \]

X and Y are independent iff:

\[ P(X, Y) = P(X) P(Y) \]

\[ P(X | Y) = P(X) \]

Rules of Probability

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: Marginalisation of the joint probability over Y.

LHS: Marginal probability of X.

Application:

\[ P(X) = \sum_{Y} P(X | Y) P(Y) \]

Example: determining the sex of fish

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

Relative frequencies of fish length

0 5 10 15 20

Length / cm

0 20 40

Frequency

Lengths of male fish

Lengths of female fish

(NB: different example from the one in Note 5.)
Example: determining the sex of fish

Fish questions

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

Relative frequency of male fish length: \( P(x | \text{M}) \)
Relative frequency of female fish length: \( P(x | \text{F}) \)

How to classify 10 cm fish?

If \( P(x | \text{M}) > P(x | \text{F}) \) ⇒ male fish
If \( P(x | \text{M}) < P(x | \text{F}) \) ⇒ female fish

Bayes' Theorem

\[
P(H | E) = \frac{P(E | H) P(H)}{P(E)}
\]

Thomas Bayes (?) (1702? – 1761)

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

Bayes' paper: http://www.jstor.org/stable/106741
http://dx.doi.org/10.1093/biomet/45.3-4.296

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

MacKay textbook, amongst many others
Inferring labels for $x = 11$ (cont.)

Equal prior probabilities:

$P(M | x = 11) = P(x = 11 | M) P(M) / P(x = 11 | F) P(F) = 0.14 \cdot 0.5 / 0.10 \cdot 2/3 = 1.4$

Classify it as male:

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

$P(M | x = 11) = P(x = 11 | M) P(M) / P(x = 11 | F) P(F) = 0.14 \cdot 1/3 / 0.10 \cdot 2/3 = 0.7$

Classify it as female

Likelihood vs posterior probability

$P(C_k | x) = P(x | C_k) P(C_k) / \sum_{k=1}^{K} P(x | C_k) P(C_k)$

Likelihood vs posterior probability (cont.)

$P(C_k | x) = P(x | C_k) P(C_k) / \sum_{k=1}^{K} P(x | C_k) P(C_k)$

Some more questions

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

1. What is the value of $P(M | X = 4)$?
2. What is the value of $P(F | X = 18)$?
3. You observe data point $x = 22$. To which class should it be assigned?
4. 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

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

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

Chain rule

$P(X_1, X_2, \ldots, X_K) = P(X_1) P(X_2 | X_1) P(X_3 | X_2, X_1) \ldots P(X_K | X_1, \ldots, X_{K-1})$

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.

  \[ \rho(X, Y) = \begin{cases} 1 & \text{if } X \text{ and } Y \text{ are perfectly positively correlated} \\ -1 & \text{if } X \text{ and } Y \text{ are perfectly negatively correlated} \\ 0 & \text{otherwise} \end{cases} \]

  \[ \rho_{XY} = \rho(Y | X) = \rho(X | Y) \]

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

  \[ P(X = -1 | Y = 0) = \frac{1}{4}, P(Y = 0 | X = -1) = \frac{1}{4} \]

  \[ P(X = 0 | Y = -1) = \frac{1}{4}, P(Y = -1 | X = 0) = \frac{1}{4} \]

  \[ P(X = 1 | Y = 1) = \frac{1}{4}, P(Y = 1 | X = 1) = \frac{1}{4} \]

  \[ \rho_{XY} = 0, \text{ but } P(X, Y) \neq P(X) P(Y) \]

  i.e., not independent

Optimisation problems we've seen

- Bayes decision rule (MAP decision rule)
  
  $k_{max} = \arg \max_{k \in C} P(C_k | x)$

- K-NN classification
  
  $c(z) = \arg \max_{j \in C} \sum_{k \in U_i(z)} \delta_{x_k}$

  where $U_i(z)$ is the set of $k$ nearest training examples to $z$.

- K-means clustering
  
  $\min E (\{m_n\})$

  $E = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{K} x_n | x_n - m_n |$

  Subject to $|u| = 1, |v| = 1, u \bot v$

- Dimensionality reduction to 2D with PCA

  Max $\Var(y) + \Var(z)$

  Subject to $|u| = 1, |v| = 1, u \bot v$

  Types of optimisation problems

- Continuous vs Discrete optimisation
- Unconstrained vs Constrained optimisation

https://en.wikipedia.org/wiki/Optimization_tree
Continuous & unconstrained optimisation problems

Optimisation of a quadratic function of one variable

Optimisation of a quadratic function of two variables

Minimisation of objective function

\[
\min_x f(x) \quad \text{where} \quad x \in \mathbb{R}^D, \quad f : \mathbb{R}^D \rightarrow \mathbb{R}
\]

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

\[
\frac{\partial f(x)}{\partial x_i} = 0, \quad \text{for} \quad i = 1, \ldots, D
\]

Vector representation:

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

\[\text{Note 5.}\]

\( ^1 \) This is not a sufficient condition, but a necessary condition.

Least square error line fitting

Optimisation problem:

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

\[
\hat{y}_n = ax_n + b
\]

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

\[
\frac{\partial E}{\partial b} = \frac{2}{N} \sum_{n=1}^{N} (ax_n + b - y_n) = 0
\]

\[
\Rightarrow \text{See the lecture note for details.}
\]

Exercise:

Optimisation problem

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

\[
\hat{y}_n = ax_n + b
\]

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

\[
\frac{\partial E}{\partial b} = \frac{2}{N} \sum_{n=1}^{N} (ax_n + b - y_n) = 0
\]

\[
\Rightarrow \text{See the lecture note for details.}
\]

Gradient descent

\[
x_{t+1} = x_t - \eta \nabla f(x)|_{x=x_t}
\]

where \( \eta > 0 \)

Things to consider:

- Choice of \( \eta \) (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’
- Pattern recognition as optimisation problem
- Most of optimisation problem does not have a
- Closed-form solution \( \Rightarrow \text{Iterative optimisation method} \)
- Check the examples in slides, and try the exercises in
- Note 5.

Mid-course feedback

Your Learn course webpage
\( \Rightarrow \text{(on the left black tab) Have Your Say} \)
\( \Rightarrow \text{Mid-course feedback} \)