F20DL and F21DL: Part 2, Machine Learning Lecture 1. Bayes rule and Bayesian learning

Katya Komendantskaya

About me

[2004 -2007] PhD from University College Cork, Ireland (Neural nets among key topics)

[2007 - 2008] Postdoctoral Year in INRIA Sophia Antipois, France

[2008 - 2010] Researcher at St Andrews University

[2010 - 2016] Lecturer, Senior Lecturer, Reader at Dundee University
(Various AI and machine learning courses, Undergrad and MSc level)

[2016 – now] Associate Professor at Heriot-Watt

How to find me

- Room G26, most days. Email to be sure.
- Open hour: Friday 13.15 - 14.15;
- Our labs: Thursdays at 10.15 and 15.15;
- Email: `ek19@hw.ac.uk`
- Phone: 0131 451 8283
- URL: `http://www.macs.hw.ac.uk/~ek19/`

Data Mining course – Part 2

- You have learned how to process data (prepare, clean, analyse)

In the 2nd part,

we will look at some most famous machine learning algorithms: supervised and unsupervised learning, Bayesian learning and Clustering, Neural nets and Decision trees

- How are they defined, as algorithms?
- How to understand what they do?
- How they work?
- How to use them for data mining?

Remaining Coursework

CW	What about	When to start?

Remaining Coursework

CW	What about	When to start?
Test 1	Bayesian Learning	Now, Week 6

Remaining Coursework

CW	What about	When to start?
Test 1	Bayesian Learning	Now, Week 6
Test 2	Clustering	Week 7

Remaining Coursework

CW	What about	When to start?
Test 1	Bayesian Learning	Now, Week 6
Test 2	Clustering	Week 7
CW2	Unsupervised learning: Bayesian Learning and Clustering ("Reading week" on Week 8)	Weeks 7-8, after Test 1 and 2

Remaining Coursework

CW	What about	When to start?
Test 1	Bayesian Learning	Now, Week 6
Test 2	Clustering	Week 7
CW2	Unsupervised learning: Bayesian Learning and Clustering ("Reading week" on Week 8)	Weeks 7-8, after Test 1 and 2
Test 3	Decision trees and Regression	Week 9

Remaining Coursework

CW	What about	When to start?
Test 1	Bayesian Learning	Now, Week 6
Test 2	Clustering	Week 7
CW2	Unsupervised learning: Bayesian Learning and Clustering ("Reading week" on Week 8)	Weeks 7-8, after Test 1 and 2
Test 3	Decision trees and Regression	Week 9
Test 4	Neural Nets	Week 10

Remaining Coursework

CW	What about	When to start?
Test 1	Bayesian Learning	Now, Week 6
Test 2	Clustering	Week 7
CW2	Unsupervised learning: Bayesian Learning and Clustering (“Reading week” on Week 8)	Weeks 7-8, after Test 1 and 2
Test 3	Decision trees and Regression	Week 9
Test 4	Neural Nets	Week 10
CW3	Supervised Learning: Decision trees and Neural Networks (“Revision week” on Week 11)	Weeks 9-11, after Test 2 and 3

Remaining Coursework

CW	What about	When to start?
Test 1	Bayesian Learning	Now, Week 6
Test 2	Clustering	Week 7
CW2	Unsupervised learning: Bayesian Learning and Clustering ("Reading week" on Week 8)	Weeks 7-8, after Test 1 and 2
Test 3	Decision trees and Regression	Week 9
Test 4	Neural Nets	Week 10
CW3	Supervised Learning: Decision trees and Neural Networks ("Revision week" on Week 11)	Weeks 9-11, after Test 2 and 3

CW2 and CW3 are group work (same groups), but tests are individually submitted and marked. Passing two tests is a pre-requisite for getting your full CW2/CW3 mark.

Coursework 2:

A real, industrial-size, data set

Facial Emotion Recognition from CW1

- You will be asked to (creatively) use algorithms we consider in the lectures on this data set
- CW2 will rely on first two weeks of lectures, with deadline on the 7th of November.
- Already "today": check out the CW spec on Vision:
  - distribute jobs and tasks as per my schedule
  - decide: Weka GUI or command line? embedded into your favourite language (Java, Bash, ...)?

Later today, we will see a simplified example of emotion recognition set.

Tests

- General pattern: During the lectures on Thursday and Friday, I announce and explain the test questions arising from the lecture materials

Tests

- General pattern: During the lectures on Thursday and Friday, I announce and explain the test questions arising from the lecture materials
- You start thinking about the questions, solve them and submit the solutions as a test on Vision at your earliest convenience, but no later than the deadline.

Tests

- General pattern: During the lectures on Thursday and Friday, I announce and explain the test questions arising from the lecture materials
- You start thinking about the questions, solve them and submit the solutions as a test on Vision at your earliest convenience, but no later than the deadline.
- The test automatically closes at 10.00 on Friday that follows the lectures it tests.
- Tests also serve us as lab exercises – so you will be able to ask questions during the lab, and you will have 3 attempts at each test.

Tests

- General pattern: During the lectures on Thursday and Friday, I announce and explain the test questions arising from the lecture materials
- You start thinking about the questions, solve them and submit the solutions as a test on Vision at your earliest convenience, but no later than the deadline.
- The test automatically closes at 10.00 on Friday that follows the lectures it tests.
- Tests also serve us as lab exercises – so you will be able to ask questions during the lab, and you will have 3 attempts at each test.
- When you submit the test, you get your mark immediately, but exact answers will be available only after the deadline.

Tests

- General pattern: During the lectures on Thursday and Friday, I announce and explain the test questions arising from the lecture materials
- You start thinking about the questions, solve them and submit the solutions as a test on Vision at your earliest convenience, but no later than the deadline.
- The test automatically closes at 10.00 on Friday that follows the lectures it tests.
- Tests also serve us as lab exercises – so you will be able to ask questions during the lab, and you will have 3 attempts at each test.
- When you submit the test, you get your mark immediately, but exact answers will be available only after the deadline.
- Tests are for fun, and to help you to better understand the lectures and the algorithms you use in CW2 and CW3.

Role of Multiple-Choice testing in this course

MCQs are now a common practice in on-line and in-person job interviews

Training you in understanding and passing MCQ

is one of the pedagogical goals of this course

- 4 MCQ tests are pre-requisites to CW2 and CW3
- Exam is in MCQ format
- For the 4 lab tests – you will have 3 attempts before the deadline, and unlimited attempts after
- This is to give you plenty of time to practice and see how they work
- Lab helpers and I will be in the labs, ready to answer any questions

Plan for this week

This week, I'll give you two "practical" lectures on Probabilities, Bayes rules and Bayes nets.

My goals are:

- Give you "simple enough" material so that you can understand every little detail as "your own".

Plan for this week

This week, I'll give you two "practical" lectures on Probabilities, Bayes rules and Bayes nets.

My goals are:

- Give you "simple enough" material so that you can understand every little detail as "your own".
- Give you a lot of practice - hence we will have practical tasks at the end of every lecture and then tests

Plan for this week

This week, I'll give you two "practical" lectures on Probabilities, Bayes rules and Bayes nets.

My goals are:

- Give you "simple enough" material so that you can understand every little detail as "your own".
- Give you a lot of practice - hence we will have practical tasks at the end of every lecture and then tests
- You will go on to use methods similar to the ones you see this week, but in a bigger scale, faster software, of real-life value

Plan for this week

This week, I'll give you two "practical" lectures on Probabilities, Bayes rules and Bayes nets.

My goals are:

- Give you "simple enough" material so that you can understand every little detail as "your own".
- Give you a lot of practice - hence we will have practical tasks at the end of every lecture and then tests
- You will go on to use methods similar to the ones you see this week, but in a bigger scale, faster software, of real-life value
- When you use more sophisticated tools, I would hope that your clear knowledge of "simple things" will support you, and help you to have a firm ground when you need to tackle harder problems

What probabilities are about

- Like logical assertions, probabilistic assertions are about possible worlds.

What probabilities are about

- Like logical assertions, probabilistic assertions are about possible worlds.
- Logical assertions characterise events as true or false, probabilistic assertions talk about how probable the various events are.

What probabilities are about

- Like logical assertions, probabilistic assertions are about possible worlds.
- Logical assertions characterise events as true or false, probabilistic assertions talk about how probable the various events are.
- The set of all possible worlds is called a sample space.
- The possible worlds are mutually exclusive and exhaustive.

What probabilities are about

- Like logical assertions, probabilistic assertions are about possible worlds.
- Logical assertions characterise events as true or false, probabilistic assertions talk about how probable the various events are.
- The set of all possible worlds is called a sample space.
- The possible worlds are mutually exclusive and exhaustive.

Example

Rolling two dice, there are 36 possible worlds to consider:
$\Omega = \{(1,1),(1,2),...,(6,6)\}$, with one world e.g. $\omega = (1,1)$.

Probability model/measure

- is a function from sets of worlds into positive real numbers that associates probability $P(\omega)$ with each possible world.

Two basic laws:

$0 \leq P(\omega) \leq 1$
$\sum P(\omega) = 1$

Probability model/measure

- is a function from sets of worlds into positive real numbers that associates probability $P(\omega)$ with each possible world.

Two basic laws:

$0 \leq P(\omega) \leq 1$
$\sum P(\omega) = 1$

Example

Probability of each one of the 36 events is $\frac{1}{36}$.
(When all events are equally likely)

Customer habits:

- A customer normally buys 10 different items every month at M&S, Tesco, and Amazon.
- M&S: brand toiletries, one Belgian chocolate box, a bottle of wine, a children's book, flowers.
- Tesco: toothpaste, soap, coffee, a gift voucher.
- Amazon: a kindle book.

Question: if we know that the customer has just made some purchase in one of these shops, what is the probability that the person bought gift vouchers?

Customer habits:

- A customer normally buys 10 different items every month at M&S, Tesco, and Amazon.
- M&S: brand toiletries, one Belgian chocolate box, a bottle of wine, a children's book, flowers.
- Tesco: toothpaste, soap, coffee, a gift voucher.
- Amazon: a kindle book.

Question: if we know that the customer has just made some purchase in one of these shops, what is the probability that the person bought gift vouchers?

Answer: $\frac{1}{10} = 0{,}1$.

Possible worlds and events

- Probabilistic assertions are not usually about particular possible worlds, but about sets of them.

Possible worlds and events

- Probabilistic assertions are not usually about particular possible worlds, but about sets of them.
- For example, we might be interested in the cases when the two dice add up to 11, the cases where doubles are rolled, etc. In probability theory, these sets are called events.

Possible worlds and events

- Probabilistic assertions are not usually about particular possible worlds, but about sets of them.
- For example, we might be interested in the cases when the two dice add up to 11, the cases where doubles are rolled, etc. In probability theory, these sets are called events.
- They are also described by *propositions* in a formal language.

Possible worlds and events

- Probabilistic assertions are not usually about particular possible worlds, but about sets of them.
- For example, we might be interested in the cases when the two dice add up to 11, the cases where doubles are rolled, etc. In probability theory, these sets are called events.
- They are also described by *propositions* in a formal language.
- For each proposition, the corresponding set contains just those possible words in which proposition holds.

Probability of propositions:

- The probability associated with a proposition is defined to be the sum of the probabilities of the worlds in which it holds:

For any proposition q, $P(q) = \sum_{\omega \in \boldsymbol{q}} P(\omega)$.

Probability of propositions:

- The probability associated with a proposition is defined to be the sum of the probabilities of the worlds in which it holds:

For any proposition q, $P(q) = \sum_{\omega \in \boldsymbol{q}} P(\omega)$.

Example

When rolling fair dice, we have
$P(Total = 11) = P(5, 6) + P(6, 5) = \frac{1}{36} + \frac{1}{36} = \frac{2}{36} = \frac{1}{18}$.
$P(doubles) = \ldots$

Probability of propositions:

- The probability associated with a proposition is defined to be the sum of the probabilities of the worlds in which it holds:

For any proposition q, $P(q) = \sum_{\omega \in \boldsymbol{q}} P(\omega)$.

Example

When rolling fair dice, we have
$P(Total = 11) = P(5, 6) + P(6, 5) = \frac{1}{36} + \frac{1}{36} = \frac{2}{36} = \frac{1}{18}$.
$P(doubles) = ... \frac{1}{6}$.

Customer Habits:

- A customer normally buys 10 different items every month at M&S, Tesco, and Amazon.
- M&S: brand toiletries, one Belgian chocolate box, a bottle of wine, a children's book, flowers.
- Tesco: toothpaste, soap, coffee, a gift voucher.
- Amazon: a kindle book.

Question: for any random transaction with these retailers this month, what is the probability that the customer buys food? Anything from M&S?

Customer Habits:

- A customer normally buys 10 different items every month at M&S, Tesco, and Amazon.
- M&S: brand toiletries, one Belgian chocolate box, a bottle of wine, a children's book, flowers.
- Tesco: toothpaste, soap, coffee, a gift voucher.
- Amazon: a kindle book.

Question: for any random transaction with these retailers this month, what is the probability that the customer buys food? Anything from M&S?

Answer: $P(food) = P(chocolate) + P(wine) + P(coffee) =$

Customer Habits:

- A customer normally buys 10 different items every month at M&S, Tesco, and Amazon.
- M&S: brand toiletries, one Belgian chocolate box, a bottle of wine, a children's book, flowers.
- Tesco: toothpaste, soap, coffee, a gift voucher.
- Amazon: a kindle book.

Question: for any random transaction with these retailers this month, what is the probability that the customer buys food? Anything from M&S?

Answer: $P(food) = P(chocolate) + P(wine) + P(coffee) = 0{,}3.$

Customer Habits:

- A customer normally buys 10 different items every month at M&S, Tesco, and Amazon.
- M&S: brand toiletries, one Belgian chocolate box, a bottle of wine, a children's book, flowers.
- Tesco: toothpaste, soap, coffee, a gift voucher.
- Amazon: a kindle book.

Question: for any random transaction with these retailers this month, what is the probability that the customer buys food? Anything from M&S?

Answer: $P(food) = P(chocolate) + P(wine) + P(coffee) = 0{,}3$.
$P(M\&S) =$
$P(toiletries) + P(chocolate) + P(wine) + P(books) + P(flowers) =$

Customer Habits:

- A customer normally buys 10 different items every month at M&S, Tesco, and Amazon.
- M&S: brand toiletries, one Belgian chocolate box, a bottle of wine, a children's book, flowers.
- Tesco: toothpaste, soap, coffee, a gift voucher.
- Amazon: a kindle book.

Question: for any random transaction with these retailers this month, what is the probability that the customer buys food? Anything from M&S?

Answer: $P(food) = P(chocolate) + P(wine) + P(coffee) = 0{,}3$.
$P(M\&S) = P(toiletries) + P(chocolate) + P(wine) + P(books) + P(flowers) = 0{,}5$.
(Remember these numbers, there will be a quiz soon)

Connection to your previous block of lectures:

We have said: For any proposition q,
$P(q = true) = \sum_{\omega \in q=true} P(\omega)$.

Connection to your previous block of lectures:

We have said: For any proposition q,
$P(q = true) = \sum_{\omega \in q = true} P(\omega)$.

Generalisation to real-valued domains

For a random variable X,
$P(a \leq X \leq b) = \int_a^b p(X) dX$
where
$p(X)$ is a probability distribution function

Connection to your previous block of lectures:

We have said: For any proposition q,
$P(q = true) = \sum_{\omega \in q = true} P(\omega)$.

Generalisation to real-valued domains

For a random variable X,
$P(a \leq X \leq b) = \int_a^b p(X)dX$
where
$p(X)$ is a probability distribution function

You considered several such distribution functions this term

For example, Gaussian or normal distribution is:

$$p(X) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}((X-\mu)/\sigma)^2}$$

where μ is mean and σ is the standard deviation.

Unconditional Probabilities

- Probabilities such as $P(Total = 11)$ and $P(food)$ are called unconditional probabilities. They refer to degrees of belief in propositions in absence of any other information.
- Most of the time, we have some additional information, called evidence...

Conditional Probabilities

Example

- E.g., the first die is already showing 5, and we are waiting for the second die.
- In this case we are interested in the conditional probability of rolling doubles given that the first die is a 5.
- This probability is written

$$P(Doubles|Die1 = 5)$$

Conditional Probability - Formula

- For any propositions a and b,

$$P(a|b) = \frac{P(a \wedge b)}{P(b)},$$

Which holds whenever $P(b) > 0$.

Conditional Probability - Formula

- For any propositions a and b,

$$P(a|b) = \frac{P(a \wedge b)}{P(b)},$$

Which holds whenever $P(b) > 0$.

Example

$$P(Doubles|Die1 = 5) = \frac{P(Doubles \wedge Die1{=}5)}{P(Die1{=}5)} = \frac{\frac{1}{36}}{1/6} = \frac{1}{6}.$$

Customer Habits:

- A customer normally buys 10 different items every month at M&S, Tesco, and Amazon.
- M&S: brand toiletries, one Belgian chocolate box, a bottle of wine, a children's book, flowers.
- Tesco: toothpaste, soap, coffee, a gift voucher.
- Amazon: a kindle book.

Question: what is the probability that the customer buys food if we know he is now buying from M&S?

Customer Habits:

- A customer normally buys 10 different items every month at M&S, Tesco, and Amazon.
- M&S: brand toiletries, one Belgian chocolate box, a bottle of wine, a children's book, flowers.
- Tesco: toothpaste, soap, coffee, a gift voucher.
- Amazon: a kindle book.

Question: what is the probability that the customer buys food if we know he is now buying from M&S?

Answer: $P(food|M\&S) = \frac{P(food \wedge (M\&S))}{P(M\&S)}$

Customer Habits:

- A customer normally buys 10 different items every month at M&S, Tesco, and Amazon.
- M&S: brand toiletries, one Belgian chocolate box, a bottle of wine, a children's book, flowers.
- Tesco: toothpaste, soap, coffee, a gift voucher.
- Amazon: a kindle book.

Question: what is the probability that the customer buys food if we know he is now buying from M&S?

Answer: $P(food|M\&S) = \frac{P(food \wedge (M\&S))}{P(M\&S)} = \frac{0{,}2}{0{,}5} = 0{,}4.$

Stop and think: Intuitive result?

Customer Habits:

- A customer normally buys 10 different items every month at M&S, Tesco, and Amazon.
- M&S: brand toiletries, one Belgian chocolate box, a bottle of wine, a children's book, flowers.
- Tesco: toothpaste, soap, coffee, a gift voucher.
- Amazon: a kindle book.

Question: what is the probability that the customer buys food if we know he is now buying from M&S?

Answer: $P(food|M\&S) = \frac{P(food \wedge (M\&S))}{P(M\&S)} = \frac{0{,}2}{0{,}5} = 0{,}4.$

Stop and think: Intuitive result?

(remember this number too)

Other probability rules

$$P(\neg a) = 1 - P(a)$$
$$P(a \vee b) = P(a) + P(b) - P(a \wedge b)$$

Other probability rules

$P(\neg a) = 1 - P(a)$
$P(a \vee b) = P(a) + P(b) - P(a \wedge b)$

To prove two events are independent, One should show that

$P(a \wedge b) = P(a)P(b)$
or
$P(a|b) = P(a)$
The former is also known as the product rule.
It will play a very important role when we consider the Bayes nets tomorrow

Probability: Frequentist vs. Bayesian

Probability: Frequentist vs. Bayesian

1. Frequentist view: probabilities from observations

- probability of heads = No of heads / No of flips
- probability of heads this time = probability of heads (history)

Probability: Frequentist vs. Bayesian

1. Frequentist view: probabilities from observations

- probability of heads = No of heads / No of flips
- probability of heads this time = probability of heads (history)

2. The objectivist view: probabilities are real aspects of the universe.

Uncertainty is ontological: pertaining to the world.

Probability: Frequentist vs. Bayesian

1. Frequentist view: probabilities from observations

- probability of heads = No of heads / No of flips
- probability of heads this time = probability of heads (history)

2. The objectivist view: probabilities are real aspects of the universe.

Uncertainty is ontological: pertaining to the world.

3. Bayesian view (subjectivist):

- probability of heads this time = agent's belief about this event (and beliefs may change!)
- belief of agent A is based on previous experience of agent A (experience changes, too)
- Uncertainty is epistemological: pertaining to the knowledge

From Conditional to Bayesian probability

- The conditional probability $P(a|b) = \frac{P(a \wedge b)}{p(b)}$ can be re-expressed as
- $P(a \wedge b) = P(a|b)P(b)$ or
- $P(b \wedge a) = P(b|a)P(a)$

From Conditional to Bayesian probability

- The conditional probability $P(a|b) = \frac{P(a \wedge b)}{p(b)}$ can be re-expressed as
- $P(a \wedge b) = P(a|b)P(b)$ or
- $P(b \wedge a) = P(b|a)P(a)$
  We now rely on the fact that $P(a \wedge b) = P(b \wedge a)$,

and so,...

$P(a|b)P(b) = P(b|a)P(a)$

Bayes' rule

Taking the formula

$P(a|b)P(b) = P(b|a)P(a)$
... and dividing both sides by $P(a)$,

- We obtain Bayes' rule: $P(b|a) = \frac{P(a|b)P(b)}{P(a)}$
- The rule underlies most of modern AI systems. Often, we perceive as evidence the effect of some unknown cause, and would like to determine the cause.

Bayes' rule

Taking the formula

$P(a|b)P(b) = P(b|a)P(a)$
... and dividing both sides by $P(a)$,

- We obtain Bayes' rule: $P(b|a) = \frac{P(a|b)P(b)}{P(a)}$
- The rule underlies most of modern AI systems. Often, we perceive as evidence the effect of some unknown cause, and would like to determine the cause.

$P(cause|effect)$ – describes diagnostic relation:

$P(cause|effect) = \frac{P(effect|cause)P(cause)}{P(effect)}$

Bayes' rule

Taking the formula

$P(a|b)P(b) = P(b|a)P(a)$
... and dividing both sides by $P(a)$,

- We obtain Bayes' rule: $P(b|a) = \frac{P(a|b)P(b)}{P(a)}$
- The rule underlies most of modern AI systems. Often, we perceive as evidence the effect of some unknown cause, and would like to determine the cause.

$P(cause|effect)$ – describes diagnostic relation:

$P(cause|effect) = \frac{P(effect|cause)P(cause)}{P(effect)}$

Note: knowing $P(effect|cause)$ is a matter of routine observation, but finding out $P(cause|effect)$ – amounts to learning, acquiring new knowledge about the world!

Example: diagnostics

- Doctor knows $P(symptoms|disease)$ and determines $P(disease|symptoms)$.

Example: diagnostics

- Doctor knows $P(symptoms|disease)$ and determines $P(disease|symptoms)$.

Example

$P(meningitis\ causes\ stiff\ \ neck) = 70\,\%$ (Written as $P(stiff\ \ neck|meningitis) = 0{,}7$)

Example: diagnostics

- Doctor knows $P(symptoms|disease)$ and determines $P(disease|symptoms)$.

Example

$P(meningitis\ causes\ stiff\ \ neck) = 70\,\%$ (Written as $P(stiff\ \ neck|meningitis) = 0{,}7$)
$P(unconditional\ \ meningitis) = \frac{1}{50000} = 0{,}00002$

Example: diagnostics

- Doctor knows $P(symptoms|disease)$ and determines $P(disease|symptoms)$.

Example

$P(meningitis\ causes\ stiff\ \ neck) = 70\,\%$ (Written as $P(stiff\ \ neck|meningitis) = 0{,}7$)
$P(unconditional\ \ meningitis) = \frac{1}{50000} = 0{,}00002$
$P(any\ \ patient\ \ has\ a\ stiff\ \ neck) = 1\,\%$ or $= 0{,}01$.

These are all simply taken from the history of medical records (no need to be a doctor to know).

Example: diagnostics

- Doctor knows $P(symptoms|disease)$ and determines $P(disease|symptoms)$.

Example

$P(meningitis\ causes\ stiff\ \ neck) = 70\,\%$ (Written as $P(stiff\ \ neck|meningitis) = 0{,}7$)
$P(unconditional\ \ meningitis) = \frac{1}{50000} = 0{,}00002$
$P(any\ \ patient\ \ has\ \ a\ \ stiff\ \ neck) = 1\,\%$ or $= 0{,}01$.

These are all simply taken from the history of medical records (no need to be a doctor to know).

$P(meningitis|stiff\ \ neck) = \frac{P(stiff\ \ neck|meningitis)P(meningitis)}{P(stiff\ \ neck)} = (0{,}7 * 0{,}00002)/0{,}01 = 0{,}0014$

Note: there has been revision of knowledge about patient's health – from negligible 0,00002 to a somewhat higher 0,0014.

Example: diagnostics

- Doctor knows $P(symptoms|disease)$ and determines $P(disease|symptoms)$.

Example

$P(meningitis\ causes\ stiff\ \ neck) = 70\,\%$ (Written as $P(stiff\ \ neck|meningitis) = 0{,}7$)
$P(unconditional\ \ meningitis) = \frac{1}{50000} = 0{,}00002$
$P(any\ \ patient\ \ has\ \ a\ \ stiff\ \ neck) = 1\,\%$ or $= 0{,}01$.

These are all simply taken from the history of medical records (no need to be a doctor to know).

$P(meningitis|stiff\ \ neck) = \frac{P(stiff\ \ neck|meningitis)P(meningitis)}{P(stiff\ \ neck)} = (0{,}7 * 0{,}00002)/0{,}01 = 0{,}0014$

Note: there has been revision of knowledge about patient's health – from negligible 0,00002 to a somewhat higher 0,0014.

Basis for many on-line "intelligent" diagnostic tools (e.g. at Boots, NHS)

Customer preferences:

- A customer normally buys 10 different items every month at M&S, Tesco, and Amazon.
- M&S: brand toiletries, one Belgian chocolate box, a bottle of wine, a children's book, flowers.
- Tesco: toothpaste, soap, coffee, a gift voucher.
- Amazon: a kindle book.

Question: if the customer bought food [from any of the shops we monitor], what is the probability that he bought it from M&S?

Customer preferences:

- A customer normally buys 10 different items every month at M&S, Tesco, and Amazon.
- M&S: brand toiletries, one Belgian chocolate box, a bottle of wine, a children's book, flowers.
- Tesco: toothpaste, soap, coffee, a gift voucher.
- Amazon: a kindle book.

Question: if the customer bought food [from any of the shops we monitor], what is the probability that he bought it from M&S?

Answer: using Bayes' law:

$$P(M\&S|food) = \frac{P(food|M\&S) * P(M\&S)}{P(food)} =$$

Customer preferences:

- A customer normally buys 10 different items every month at M&S, Tesco, and Amazon.
- M&S: brand toiletries, one Belgian chocolate box, a bottle of wine, a children's book, flowers.
- Tesco: toothpaste, soap, coffee, a gift voucher.
- Amazon: a kindle book.

Question: if the customer bought food [from any of the shops we monitor], what is the probability that he bought it from M&S?

Answer: using Bayes' law:
$P(M\&S|food) = \frac{P(food|M\&S)*P(M\&S)}{P(food)} =$
by previous calculations of the components
$= 0{,}4 * 0{,}5/0{,}3 \approx 0{,}67$.

Stop and think: What has just happened?

- We have learned from experience (= revised our beliefs!):

Our Prior (default) knowledge about the customer's habits was: $P(M\&S) = 0{,}5$

We have just substantially revised our default belief in probability of M&S purchase. Now (after the observation), it is 67 %!

Using Bayes' law, every new observation will lead to knowledge revision!

and now...

Apply Bayesian learning in data mining!

Internet advertisements

- An internet shop wants to have an “intelligent” program that generates tailored advertisements for each customer.
- A chosen customer has the history of the following 10 actions with the shopping basket:

Customer transactions

Trans.	Music on CD?	Music on MP3?	Board Games	On-line Games	Output
T1	No	Yes	No	Yes	Buys
T2	Yes	No	No	No	Cancels
T3	Yes	No	No	Yes	Buys
T4	Yes	No	Yes	No	Cancels
T5	No	Yes	No	No	Cancels
T6	No	Yes	Yes	No	Cancels
T7	No	No	No	Yes	Buys
T8	No	Yes	Yes	Yes	Cancels
T9	Yes	Yes	No	No	Cancels
T10	Yes	Yes	No	Yes	Buys

Question 1

What is the Prior Probability $P(A)$ of the target feature A, where A is the following event

- "Customer buys the products"

Trans.	Music on CD?	Music on MP3?	Board Games	On-line Games	Output
T1	No	Yes	No	Yes	Buys
T2	Yes	No	No	No	Cancels
T3	Yes	No	No	Yes	Buys
T4	Yes	No	Yes	No	Cancels
T5	No	Yes	No	No	Cancels
T6	No	Yes	Yes	No	Cancels
T7	No	No	No	Yes	Buys
T8	No	Yes	Yes	Yes	Cancels
T9	Yes	Yes	No	No	Cancels
T10	Yes	Yes	No	Yes	Buys

Question 1

What is the Prior Probability $P(A)$ of the target feature A, where A is the following event

- "Customer buys the products"

Trans.	Music on CD?	Music on MP3?	Board Games	On-line Games	Output
T1	No	Yes	No	Yes	Buys
T2	Yes	No	No	No	Cancels
T3	Yes	No	No	Yes	Buys
T4	Yes	No	Yes	No	Cancels
T5	No	Yes	No	No	Cancels
T6	No	Yes	Yes	No	Cancels
T7	No	No	No	Yes	Buys
T8	No	Yes	Yes	Yes	Cancels
T9	Yes	Yes	No	No	Cancels
T10	Yes	Yes	No	Yes	Buys

- $P(A) = \frac{4}{10} = 0{,}4$

Question 2

Compute the Conditional Probability $P(B|A) = \frac{P(A \wedge B)}{P(A)}$ of event B given event A, where A is as above, and B is the training feature:

- "CDs are bought"

Trans.	Music on CD?	Music on MP3?	Board Games	On-line Games	Output
T1	No	Yes	No	Yes	Buys
T2	Yes	No	No	No	Cancels
T3	Yes	No	No	Yes	Buys
T4	Yes	No	Yes	No	Cancels
T5	No	Yes	No	No	Cancels
T6	No	Yes	Yes	No	Cancels
T7	No	No	No	Yes	Buys
T8	No	Yes	Yes	Yes	Cancels
T9	Yes	Yes	No	No	Cancels
T10	Yes	Yes	No	Yes	Buys

Question 2

Compute the Conditional Probability $P(B|A) = \frac{P(A \wedge B)}{P(A)}$ of event B given event A, where A is as above, and B is the training feature:

- "CDs are bought"

Trans.	Music on CD?	Music on MP3?	Board Games	On-line Games	Output
T1	No	Yes	No	Yes	Buys
T2	Yes	No	No	No	Cancels
T3	Yes	No	No	Yes	Buys
T4	Yes	No	Yes	No	Cancels
T5	No	Yes	No	No	Cancels
T6	No	Yes	Yes	No	Cancels
T7	No	No	No	Yes	Buys
T8	No	Yes	Yes	Yes	Cancels
T9	Yes	Yes	No	No	Cancels
T10	Yes	Yes	No	Yes	Buys

- $P(A \wedge B) = 0{,}2$. So, $P(B|A) = \frac{P(A \wedge B)}{P(A)} = \frac{0{,}2}{0{,}4} = 0{,}5$

Question 3

Compute the Bayesian Probability $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ of event A given event B, where A and B are as above.

Trans.	Music on CD?	Music on MP3?	Board Games	On-line Games	Output
T1	No	Yes	No	Yes	Buys
T2	Yes	No	No	No	Cancels
T3	Yes	No	No	Yes	Buys
T4	Yes	No	Yes	No	Cancels
T5	No	Yes	No	No	Cancels
T6	No	Yes	Yes	No	Cancels
T7	No	No	No	Yes	Buys
T8	No	Yes	Yes	Yes	Cancels
T9	Yes	Yes	No	No	Cancels
T10	Yes	Yes	No	Yes	Buys

Question 3

Compute the Bayesian Probability $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ of event A given event B, where A and B are as above.

Trans.	Music on CD?	Music on MP3?	Board Games	On-line Games	Output
T1	No	Yes	No	Yes	Buys
T2	Yes	No	No	No	Cancels
T3	Yes	No	No	Yes	Buys
T4	Yes	No	Yes	No	Cancels
T5	No	Yes	No	No	Cancels
T6	No	Yes	Yes	No	Cancels
T7	No	No	No	Yes	Buys
T8	No	Yes	Yes	Yes	Cancels
T9	Yes	Yes	No	No	Cancels
T10	Yes	Yes	No	Yes	Buys

- $P(B) = 0{,}5$. So, $P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{0{,}5*0{,}4}{0{,}5} = 0{,}4$

An important side note:

We have only worked with a one-feature case:

Trans.	Music on CD?				Output
T1	No				Buys
T2	Yes				Cancels
T3	Yes				Buys
T4	Yes				Cancels
T5	No				Cancels
T6	No				Cancels
T7	No				Buys
T8	No				Cancels
T9	Yes				Cancels
T10	Yes				Buys

... Will restore the full picture tomorrow!

Conclusions from the experiment

- We have practiced to work with data represented by examples/features/labels.
- It is now easy to see how Bayesian learning works on data.
- As chance has it, we did not really change the knowledge after applying the Bayes rule.
  - We started with $P(A) = 0{,}4$
  - We ended up with $P(A|B) = 0{,}4$ after the observation

  This has to do with our data. If the chosen feature is inessential, you may not get a good result out of observing it: this will certainly be true for real-life situations, not only in toy examples.

Conclusions from the experiment

- We have practiced to work with data represented by examples/features/labels.
- It is now easy to see how Bayesian learning works on data.
- As chance has it, we did not really change the knowledge after applying the Bayes rule.
  - We started with $P(A) = 0{,}4$
  - We ended up with $P(A|B) = 0{,}4$ after the observation

  This has to do with our data. If the chosen feature is inessential, you may not get a good result out of observing it: this will certainly be true for real-life situations, not only in toy examples.
- Lets play again and see if there are some important features!

Lets play with these two variables now:

Trans.	Music on MP3?	Output
T1	Yes	Buys
T2	No	Cancels
T3	No	Buys
T4	No	Cancels
T5	Yes	Cancels
T6	Yes	Cancels
T7	No	Buys
T8	Yes	Cancels
T9	Yes	Cancels
T10	Yes	Buys

Question 1

What is the Prior Probability $P(A)$ of the random variable A, where A is

- "Customer buys the products" ($P(Output = Buys)$)

Trans.	Music on MP3?	Output
T1	Yes	Buys
T2	No	Cancels
T3	No	Buys
T4	No	Cancels
T5	Yes	Cancels
T6	Yes	Cancels
T7	No	Buys
T8	Yes	Cancels
T9	Yes	Cancels
T10	Yes	Buys

Question 1

What is the Prior Probability $P(A)$ of the random variable A, where A is

- "Customer buys the products" ($P(Output = Buys)$)

Trans.	Music on MP3?	Output
T1	Yes	Buys
T2	No	Cancels
T3	No	Buys
T4	No	Cancels
T5	Yes	Cancels
T6	Yes	Cancels
T7	No	Buys
T8	Yes	Cancels
T9	Yes	Cancels
T10	Yes	Buys

- $P(A) = \frac{4}{10} = 0{,}4$

Question 2

Compute the Conditional Probability $P(B|A) = \frac{P(A \wedge B)}{P(A)}$ of variable B given A, where A as before, and B is :

- "MP3 is bought"

Trans.	Music on MP3?	Output
T1	Yes	Buys
T2	No	Cancels
T3	No	Buys
T4	No	Cancels
T5	Yes	Cancels
T6	Yes	Cancels
T7	No	Buys
T8	Yes	Cancels
T9	Yes	Cancels
T10	Yes	Buys

Question 2

Compute the Conditional Probability $P(B|A) = \frac{P(A \wedge B)}{P(A)}$ of variable B given A, where A as before, and B is :

- "MP3 is bought"

Trans.	Music on MP3?	Output
T1	Yes	Buys
T2	No	Cancels
T3	No	Buys
T4	No	Cancels
T5	Yes	Cancels
T6	Yes	Cancels
T7	No	Buys
T8	Yes	Cancels
T9	Yes	Cancels
T10	Yes	Buys

- $P(A \wedge B) = 0{,}2$. So, $P(B|A) = \frac{P(A \wedge B)}{P(A)} = \frac{0{,}2}{0{,}4} = 0{,}5$

Question 3

Compute the Bayesian Probability $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ of event A given event B, where A and B are as before.

Trans.	Music on MP3?	Output
T1	Yes	Buys
T2	No	Cancels
T3	No	Buys
T4	No	Cancels
T5	Yes	Cancels
T6	Yes	Cancels
T7	No	Buys
T8	Yes	Cancels
T9	Yes	Cancels
T10	Yes	Buys

Question 3

Compute the Bayesian Probability $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ of event A given event B, where A and B are as before.

Trans.	Music on MP3?	Output
T1	Yes	Buys
T2	No	Cancels
T3	No	Buys
T4	No	Cancels
T5	Yes	Cancels
T6	Yes	Cancels
T7	No	Buys
T8	Yes	Cancels
T9	Yes	Cancels
T10	Yes	Buys

- $P(B) = 0{,}6$. So, $P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{0{,}5*0{,}4}{0{,}6} = 0{,}33$

What has just happened?

- We have learned from experience (= revised our beliefs!):

Our Prior (default) knowledge about the customer's habits was: $P(Output = Buys) = 0{,}4$

We have just substantially revised our default belief in probability of the purchase. Now (after the observation that the customer is browsing MP3), it is 0,33!

- What does it tell us about the customer's preferences?

What has just happened?

- We have learned from experience (= revised our beliefs!):

Our Prior (default) knowledge about the customer's habits was: $P(Output = Buys) = 0{,}4$

We have just substantially revised our default belief in probability of the purchase. Now (after the observation that the customer is browsing MP3), it is 0,33!

- What does it tell us about the customer's preferences?

Bayesian Probability

is about revision of beliefs: it gives subjective, rather than objective, view on probabilities
Note: neither the customer nor the data set changed.

Conclusions from the experiment

- We have practiced to view given data in terms of possible worlds, random variables, and conditional probabilities
- It is now easy to see how Bayesian learning works on data.
- It now remains to formulate algorithms based on this initial intuition

Test 1 “toy example”: Face recognition

	X				X		
			X				
	X				X		
		X	X	X			

	X				X		
			X				
X						X	
	X				X		
		X	X	X			

	X				X		
			X				
		X	X	X			
	X				X		

An application is taught how to recognise whether a face is “happy” or “sad”. Human users have labelled some pictures for it. It needs to learn on the basis of this statistics.

Test 1 "toy example": Face recognition

	X				X		
			X				
	X				X		
		X	X	X			

	X				X		
			X				
X						X	
	X				X		
		X	X	X			

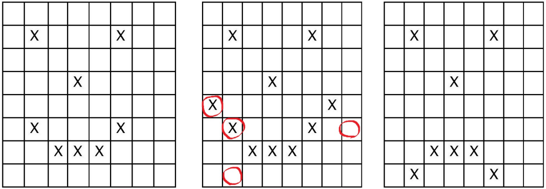
	X				X		
			X				
		X	X	X			
	X				X		

An application is taught how to recognise whether a face is "happy" or "sad". Human users have labelled some pictures for it. It needs to learn on the basis of this statistics.

I simplify our lives by clever feature extraction...

- Assume each face is symmetric, and consider only half face
- Consider only the key regions around the mouth (for a smile)

Test 1 “toy example”: Emotion recognition



I consider the left side of the face, mouth area, plus one feature of a noise (on a margin), to make it more realistic

this gives 4 features: Cells 33, 42, 48, 58

Grid face emotions

Picture	Cell 33	Cell 42	Cell 48	Cell 58	Face expression
P1	White	Black	White	White	Happy
P2	Black	Black	White	White	Happy
P3	White	White	White	Black	Sad
P4	White	White	Black	White	Sad
P5	Black	White	Black	Black	Happy
P6	White	White	Black	Black	Sad
P7	Black	White	White	Black	Sad
P8	Black	White	Black	Black	Sad
P9	White	Black	Black	Black	Sad
P10	White	Black	White	Black	Sad

Test 1 Questions (only first part)

1. What is the Prior Probability $P(A)$ of the target feature A, where A is the following event
   - "The grid face is happy"
2. Compute the Conditional Probability $P(B|A) = \frac{P(A \wedge B)}{P(A)}$ of event B given event A, where A is as above, and B is the training feature:
   - "Cell 33 is black"
3. Compute the Bayesian Probability $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ of event A given event B.
4. Was the knowledge about A revised after observation of B?
5. ...

... 10 more questions to follow tomorrow...

Conclusions

- We have just had a recap of the probability theory we will need next time;
- We have discussed some simple examples where it can be used in data-mining
- Now: test your understanding: Answer Q1-Q4
- Tomorrow: Recap of Random variables and product rule; Bayes Nets