Probability Basics

Martin Emms

September 12, 2019

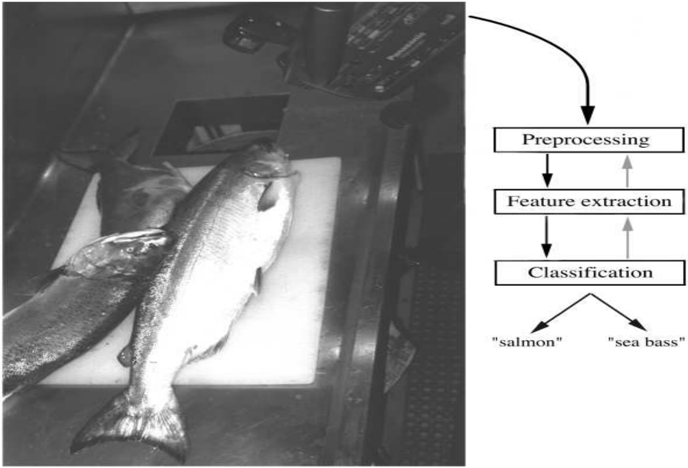
Probabilistic Inference

Probabilistic Inference

- Suppose there's a variable X whose value you would like to know, but don't
- Suppose there's another variable Y whose value you do know
- Suppose you know probabilities about how values of X and Y go together
- There's a standard way to use the probabilities to make a best guess about X
- In Speech Recognition you want to guess the words which were said, in Machine Translation you want to guess the best translation. To introduce the basic probabilistic framework we will first look though at entirely different kinds of example.

Duda and Hart's fish example

Suppose there are 2 types of fish. You might want to design a fish-sorter which seeks to distinguish between the 2 types of fish (eg. salmon vs. sea bass) by the value of some observable attribute, possibly an attribute a camera can easily measure (eg. lightness of skin)



images from Duda and Hart, Pattern Recognition

Can be formalised by representing a fish with 2 variables

- ω: a variable for the *type* of fish (values ω_1, ω_2)
- x: observed skin brightness

Can be formalised by representing a fish with 2 variables

- ω: a variable for the *type* of fish (values ω_1, ω_2)
- x: observed skin brightness

Then suppose these distributions are known:

1. $P(\omega)$
2. $P(x|\omega)$

Can be formalised by representing a fish with 2 variables

- ω: a variable for the *type* of fish (values ω_1, ω_2)
- x: observed skin brightness

Then suppose these distributions are known:

1. $P(\omega)$
2. $P(x|\omega)$

(**Jargon**: $P(x|\omega)$ might be called the *class conditional probability*)

Can be formalised by representing a fish with 2 variables

- ω: a variable for the *type* of fish (values ω_1, ω_2)
- x: observed skin brightness

Then suppose these distributions are known:

1. $P(\omega)$
2. $P(x|\omega)$

(**Jargon**: $P(x|\omega)$ might be called the *class conditional probability*)

If you observe a fish with a particular value for x, what is the best way to use the observation to predict its category?

Maximise Joint Probability

The following seems (and is) sensible

$$\textit{choose} \arg\max_{\omega} P(\omega, x)$$

i.e. pick the value for ω which together with x gives the likeliest pairing.

Maximise Joint Probability

The following seems (and is) sensible

$$\textit{choose} \arg\max_{\omega} P(\omega, x)$$

i.e. pick the value for ω which together with x gives the likeliest pairing.
Using the product rule this can be recast as

'Bayesian Classifier'

$$\textit{choose} \arg\max_{\omega} P(x|\omega)P(\omega) \tag{1}$$

Maximise Joint Probability

The following seems (and is) sensible

$$\textit{choose} \arg\max_{\omega} P(\omega, x)$$

i.e. pick the value for ω which together with x gives the likeliest pairing.
Using the product rule this can be recast as

'Bayesian Classifier'

$$\textit{choose} \arg\max_{\omega} P(x|\omega)P(\omega) \tag{1}$$

So if you know both $P(x|\omega)$ and $P(\omega)$ for the two classes ω_1 and ω_2 can now pick the one which maximises $P(x|\omega)P(\omega)$

though widely given the name 'Bayesian Classifier' this really doing nothing more than saying pick the ω which makes the combination you are looking at as likely as possible.

Maximise Conditional Probability

An equally sensible and in fact equivalent intuition for how to pick ω is to maximise conditional probability of ω given x ie.

$$\textit{choose} \arg\max_{\omega} P(\omega|x)$$

i.e. pick the value for ω which is likeliest given x.

Maximise Conditional Probability

An equally sensible and in fact equivalent intuition for how to pick ω is to maximise conditional probability of ω given x ie.

$$\textit{choose} \arg\max_{\omega} P(\omega|x)$$

i.e. pick the value for ω which is likeliest given x. This turns out to give exactly the same criterion as the maximise-joint before in (1), as follows

$$\arg\max_{\omega} P(\omega|x) \quad = \quad \arg\max_{\omega} P(\omega, x)/P(x) \tag{2}$$

Maximise Conditional Probability

An equally sensible and in fact equivalent intuition for how to pick ω is to maximise conditional probability of ω given x ie.

$$\textit{choose} \arg\max_{\omega} P(\omega|x)$$

i.e. pick the value for ω which is likeliest given x. This turns out to give exactly the same criterion as the maximise-joint before in (1), as follows

$$\arg\max_{\omega} P(\omega|x) = \arg\max_{\omega} P(\omega, x)/P(x) \quad (2)$$

$$= \arg\max_{\omega} P(x|\omega)P(\omega)/P(x) \quad (3)$$

Maximise Conditional Probability

An equally sensible and in fact equivalent intuition for how to pick ω is to maximise conditional probability of ω given x ie.

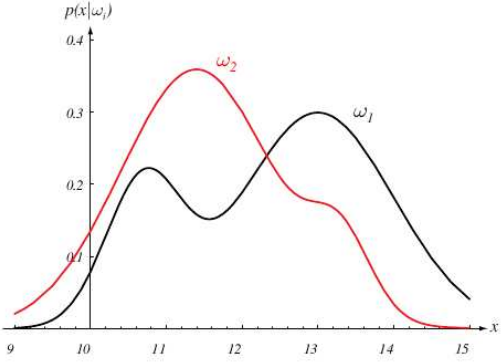
$$\textit{choose} \arg\max_{\omega} P(\omega|x)$$

i.e. pick the value for ω which is likeliest given x. This turns out to give exactly the same criterion as the maximise-joint before in (1), as follows

$$\begin{aligned} \arg\max_{\omega} P(\omega|x) &= \arg\max_{\omega} P(\omega, x)/P(x) && (2) \\ &= \arg\max_{\omega} P(x|\omega)P(\omega)/P(x) && (3) \\ &= \arg\max_{\omega} P(x|\omega)P(\omega) && (4) \end{aligned}$$

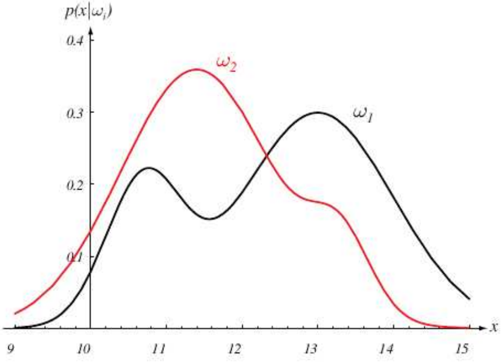
(2) is by definition of conditional probablility, (3) is by Product Rule, and (4) because denominator $P(x)$ does not mention ω, it does not vary with ω and can be left out

The following shows hypothetical plots of $P(x|\omega_1)$ and $P(x|\omega_2)$



images from Duda and Hart, Pattern Recognition

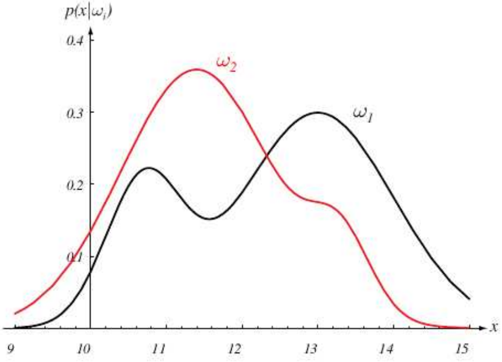
The following shows hypothetical plots of $P(x|\omega_1)$ and $P(x|\omega_2)$



images from Duda and Hart, Pattern Recognition

- Basically up to about $x = 12.5$, $P(x|\omega_2) > P(x|\omega_1)$ and thereafter the relation is the other way around.

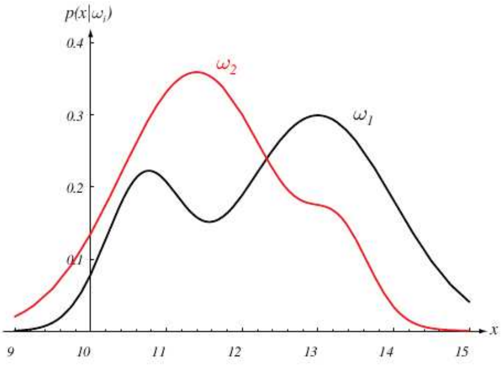
The following shows hypothetical plots of $P(x|\omega_1)$ and $P(x|\omega_2)$



images from Duda and Hart, Pattern Recognition

- Basically up to about $x = 12.5$, $P(x|\omega_2) > P(x|\omega_1)$ and thereafter the relation is the other way around.
- but this does not mean ω_2 should be chosen for $x < 12.5$, and ω_1 otherwise.

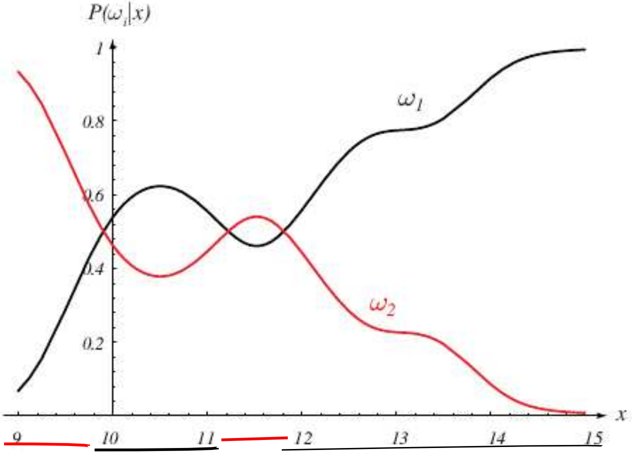
The following shows hypothetical plots of $P(x|\omega_1)$ and $P(x|\omega_2)$



images from Duda and Hart, Pattern Recognition

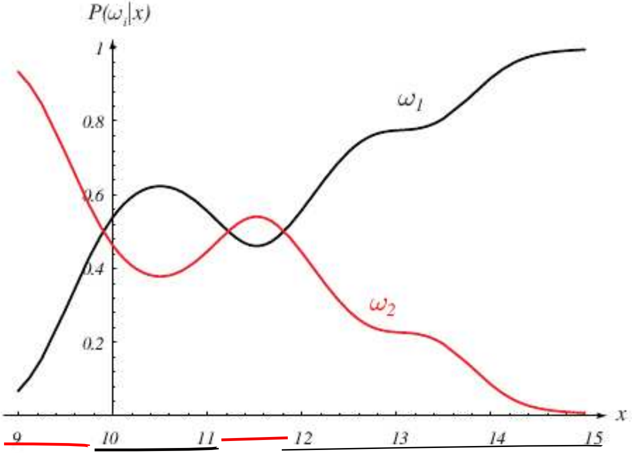
- Basically up to about $x = 12.5$, $P(x|\omega_2) > P(x|\omega_1)$ and thereafter the relation is the other way around.
- but this does not mean ω_2 should be chosen for $x < 12.5$, and ω_1 otherwise.
- the plot shows only *half* of the $P(x|\omega)P(\omega)$ referred to in the decision function (1): the other factor is the a priori likelihood $P(\omega)$

Assuming a priori probs $P(\omega_1) = 2/3$, $P(\omega_2) = 1/3$, the plots below show $P(\omega_1, x)$ and $P(\omega_2, x)$, normalised at each x by $P(x)$ (ie. it shows $P(\omega|x)$)



images from Duda and Hart, Pattern Recognition

Assuming a priori probs $P(\omega_1) = 2/3$, $P(\omega_2) = 1/3$, the plots below show $P(\omega_1, x)$ and $P(\omega_2, x)$, normalised at each x by $P(x)$ (ie. it shows $P(\omega|x)$)



images from Duda and Hart, Pattern Recognition

- So roughly for $x < 10$ or $11 < x < 12$, ω_2 is the best-guess
- So roughly for $10 < x < 11$ or $12 < x$, ω_1 is the best-guess

Optimality

Optimality

this Bayesian recipe is guaranteed to give the least error in the long term: if you know the probs $p(x|\omega)$ and $p(\omega)$, you cannot do better than always guessing $\arg\max_\omega(p(x|\omega)p(\omega))$

Optimality

this Bayesian recipe is guaranteed to give the least error in the long term: if you know the probs $p(x|\omega)$ and $p(\omega)$, you cannot do better than always guessing $\arg\max_{\omega}(p(x|\omega)p(\omega))$

there's some special cases

Optimality

this Bayesian recipe is guaranteed to give the least error in the long term: if you know the probs $p(x|\omega)$ and $p(\omega)$, you cannot do better than always guessing $\arg\max_\omega(p(x|\omega)p(\omega))$

there's some special cases

- if $p(x|\omega_1) = p(x|\omega_2)$, the evidence tells you nothing, and the decision rests entirely on $p(\omega_1)$ vs $p(\omega_2)$

Optimality

this Bayesian recipe is guaranteed to give the least error in the long term: if you know the probs $p(x|\omega)$ and $p(\omega)$, you cannot do better than always guessing $\arg\max_\omega(p(x|\omega)p(\omega))$

there's some special cases

- if $p(x|\omega_1) = p(x|\omega_2)$, the evidence tells you nothing, and the decision rests entirely on $p(\omega_1)$ vs $p(\omega_2)$
- if $p(\omega_1) = p(\omega_2)$, then the decision rests entirely on the class-conditionals: $p(x|\omega_1)$ vs. $p(x|\omega_2)$

'prior' and 'posterior'

- have seen that

$$\arg\max_{\omega} P(\omega|x) = \arg\max_{\omega}(p(x|\omega)p(\omega))$$

- often $p(\omega)$ is termed the prior probability (guessing the fish *before* looking)
- often $p(\omega|x)$ is termed the posterior probability (guessing the fish *after* looking)

'prior' and 'posterior'

- have seen that

$$\arg\max_{\omega} P(\omega|x) = \arg\max_{\omega}(p(x|\omega)p(\omega))$$

- often $p(\omega)$ is termed the prior probability (guessing the fish *before* looking)
- often $p(\omega|x)$ is termed the posterior probability (guessing the fish *after* looking)

$$\arg\max_{\omega} \underbrace{P(\omega|x)}_{\text{posterior}} = \arg\max_{\omega}(p(x|\omega)\underbrace{p(\omega)}_{\text{prior}})$$

- So can choose by considering $P(x|\omega)P(\omega)$.
- it can sometimes surprise that for all the ω, $P(x|\omega)P(\omega)$ might be tiny and not sum to one: but recall its a joint probability, so it incorporates the probability of the evidence, which might not be very likely.

- So can choose by considering $P(x|\omega)P(\omega)$.
- it can sometimes surprise that for all the ω, $P(x|\omega)P(\omega)$ might be tiny and not sum to one: but recall its a joint probability, so it incorporates the probability of the evidence, which might not be very likely.
- It also must be the case that

$$P(x) = \sum_{\omega} P(\omega, x) = \sum_{\omega} P(x|\omega)P(\omega)$$

so $P(x)$ can be obtained by summing $P(x|\omega)P(\omega)$ for the different values of ω, the same term whose maximum value is searched for in (1)

- So can choose by considering $P(x|\omega)P(\omega)$.
- it can sometimes surprise that for all the ω, $P(x|\omega)P(\omega)$ might be tiny and not sum to one: but recall its a joint probability, so it incorporates the probability of the evidence, which might not be very likely.
- It also must be the case that

$$P(x) = \sum_{\omega} P(\omega, x) = \sum_{\omega} P(x|\omega)P(\omega)$$

  so $P(x)$ can be obtained by summing $P(x|\omega)P(\omega)$ for the different values of ω, the same term whose maximum value is searched for in (1)
- so to get the true conditional $p(\omega|x)$, can get the two $P(x|\omega_1)P(\omega_1)$ and $P(x|\omega_2)P(\omega_2)$ values and then divide each by their sum

- So can choose by considering $P(x|\omega)P(\omega)$.
- it can sometimes surprise that for all the ω, $P(x|\omega)P(\omega)$ might be tiny and not sum to one: but recall its a joint probability, so it incorporates the probability of the evidence, which might not be very likely.
- It also must be the case that

$$P(x) = \sum_{\omega} P(\omega, x) = \sum_{\omega} P(x|\omega)P(\omega)$$

  so $P(x)$ can be obtained by summing $P(x|\omega)P(\omega)$ for the different values of ω, the same term whose maximum value is searched for in (1)
- so to get the true conditional $p(\omega|x)$, can get the two $P(x|\omega_1)P(\omega_1)$ and $P(x|\omega_2)P(\omega_2)$ values and then divide each by their sum
- Without dividing through by $P(x)$ you get basically the much smaller joint probabilities. The maximum occurs for the same ω as the conditional probability and the ratios amongst them are the same as amongst the conditional probs.

Jedward example

A sound-bite may or may not have been produced by JedWard. A sound-bite may or may not contain the word OMG*.*
You hear OMG *and want to work out the probability that the speaker is Jedward*

Jedward example

> *A sound-bite may or may not have been produced by JedWard. A sound-bite may or may not contain the word* OMG.
> *You hear* OMG *and want to work out the probability that the speaker is Jedward*

Formalize with 2 discrete variables

- discrete $Speaker$, values in $\{Jedward, Other\}$
- discrete OMG, values in $\{true, false\}$

Jedward example

A sound-bite may or may not have been produced by JedWard. A sound-bite may or may not contain the word OMG.
You hear OMG *and want to work out the probability that the speaker is Jedward*

Formalize with 2 discrete variables

- discrete $Speaker$, values in $\{Jedward, Other\}$
- discrete OMG, values in $\{true, false\}$

Let jed stand for $Speaker = Jedward$, omg stand for $OMG = true$
Then suppose these individual probabilities are known

1. $p(jed) = 0.01$
2. $p(omg|jed) = 1.0$
3. $p(omg|\neg jed) = 0.1$

Jedward example

> *A sound-bite may or may not have been produced by JedWard. A sound-bite may or may not contain the word* OMG.
> *You hear* OMG *and want to work out the probability that the speaker is Jedward*

Formalize with 2 discrete variables

- discrete $Speaker$, values in $\{Jedward, Other\}$
- discrete OMG, values in $\{true, false\}$

Let jed stand for $Speaker = Jedward$, omg stand for $OMG = true$
Then suppose these individual probabilities are known

1. $p(jed) = 0.01$
2. $p(omg|jed) = 1.0$
3. $p(omg|\neg jed) = 0.1$

choosing by Bayesian rule (1)

$$p(omg|jed)p(jed) = 0.01, p(omg|\neg jed)p(\neg jed) = 0.099$$

Jedward example

A sound-bite may or may not have been produced by JedWard. A sound-bite may or may not contain the word OMG.
You hear OMG *and want to work out the probability that the speaker is Jedward*

Formalize with 2 discrete variables

- discrete $Speaker$, values in $\{Jedward, Other\}$
- discrete OMG, values in $\{true, false\}$

Let jed stand for $Speaker = Jedward$, omg stand for $OMG = true$
Then suppose these individual probabilities are known

1. $p(jed) = 0.01$
2. $p(omg|jed) = 1.0$
3. $p(omg|\neg jed) = 0.1$

choosing by Bayesian rule (1)

$$p(omg|jed)p(jed) = 0.01, p(omg|\neg jed)p(\neg jed) = 0.099$$

hence choose $\neg jed$

we have

$$p(omg|jed)p(jed) = 0.01, p(omg|\neg jed)p(\neg jed) = 0.099$$

both values are quite small, and they do not sum to 1

this is because they are alternate expressions for the joint probabilities $p(omg, jed)$ and $p(omg, \neg jed)$, and summing these gives the total *omg* probability, which is not that large.

if want real probability $p(jed|omg)$ summing these alternatives and normalising by this (effectively dividing by $p(omg)$) gives

$$p(jed|omg) = 0.0917, p(\neg jed|omg) = 0.9083$$

if want real probability $p(jed|omg)$ summing these alternatives and normalising by this (effectively dividing by $p(omg)$) gives

$$p(jed|omg) = 0.0917, p(\neg jed|omg) = 0.9083$$

The posterior probability of $p(jed|omg)$ is quite small, even though $p(omg|jed)$ is large. This is due to the quite low prior $p(jed)$, and non negligible $p(omg|\neg jed) = 0.1$

if want real probability $p(jed|omg)$ summing these alternatives and normalising by this (effectively dividing by $p(omg)$) gives

$$p(jed|omg) = 0.0917, p(\neg jed|omg) = 0.9083$$

The posterior probability of $p(jed|omg)$ is quite small, even though $p(omg|jed)$ is large. This is due to the quite low prior $p(jed)$, and non negligible $p(omg|\neg jed) = 0.1$

Raising the prior prob for jed to $p(jed) = 0.1$, changes the outcome to

$$p(jed|omg) = 0.526, p(\neg jed|omg) = 0.474$$

if want real probability $p(jed|omg)$ summing these alternatives and normalising by this (effectively dividing by $p(omg)$) gives

$$p(jed|omg) = 0.0917, p(\neg jed|omg) = 0.9083$$

The posterior probability of $p(jed|omg)$ is quite small, even though $p(omg|jed)$ is large. This is due to the quite low prior $p(jed)$, and non negligible $p(omg|\neg jed) = 0.1$

Raising the prior prob for jed to $p(jed) = 0.1$, changes the outcome to

$$p(jed|omg) = 0.526, p(\neg jed|omg) = 0.474$$

Or alternatively decreasing the prob of hearing *OMG* from anyone else to 0.001, changes the outcome to

$$p(jed|omg) = 0.917, p(\neg jed|omg) = 0.083$$

Recap

Joint Probability $P(X, Y)$

Marginal Probability $P(X) = \sum_Y P(X, Y)$

Conditional Probability $P(Y|X) = \dfrac{P(X, Y)}{P(X)}$ ... really $\dfrac{count(X, Y)}{count(X)}$

Product Rule $P(X, Y) = P(Y|X) \times P(X)$

Chain Rule $P(X, Y, Z) = p(Z|(X, Y)) \times P(X, Y) =$
$p(Z|(X, Y)) \times P(Y|X) \times p(X)$

Conditional Independence $P(X|Y, Z) = P(X|Z)$ ie. X ignores Y given Z

Bayesian Inversion $P(X|Y) = \dfrac{P(Y|X)P(X)}{P(Y)}$

Inference to infer X from Y choose $X = \arg\max_X P(Y|X)P(X)$

Further reading

see the course pages under 'Course Outline' for details on particular parts of of particular books which can serve as further sources of information concerning the topics introduced by the preceding slides