Machine Learning

Anders Jonsson, **Vicenç Gómez**

Master in Intelligent Interactive Systems
2021-22

Lecture 9
Bayesian Machine Learning

Introduction

Introduction

- Two lectures on Bayesian Machine Learning
  1. Learning as Inference (Today)
  2. Inference in Probabilistic Graphical Models (next week)
- Material for Today:
  D. Mackay's book: *Information Theory, Inference, and Learning Algorithms*
  Chapters 2,3, and 28.

Introduction

- Goals of this lecture:
    - Refresh basic probability
    - Inverse probabilities vs forward probabilities
    - Learning a model as inference
    - Model comparison as inference (Occam's razor)
    - Relate this with what you learned so far

Inverse Probabilities

Recap on probability theory

Definitions:

- X is a **random variable**, takes values $x \in \mathcal{A}_X = \{a_1, a_2, \ldots, a_i, \ldots, a_I\}$ with probabilities $\mathcal{P}_X = \{p_1, p_2, \ldots, p_i, \ldots, p_I\}$

Recap on probability theory

Definitions:

- X is a **random variable**, takes values $x \in \mathcal{A}_X = \{a_1, a_2, \ldots, a_i, \ldots, a_I\}$ with probabilities $\mathcal{P}_X = \{p_1, p_2, \ldots, p_i, \ldots, p_I\}$
    - $p(x = a_i) = p_i, p_i \geq 0$

Recap on probability theory

Definitions:

- X is a **random variable**, takes values $x \in \mathcal{A}_X = \{a_1, a_2, \ldots, a_i, \ldots, a_I\}$ with probabilities $\mathcal{P}_X = \{p_1, p_2, \ldots, p_i, \ldots, p_I\}$
  - $p(x = a_i) = p_i, p_i \geq 0$
  - $\sum_{a_i \in \mathcal{A}_X} P(x = a_i) = 1$

Recap on probability theory

Definitions:

- X is a **random variable**, takes values $x \in \mathcal{A}_X = \{a_1, a_2, \dots, a_i, \dots, a_I\}$ with probabilities $\mathcal{P}_X = \{p_1, p_2, \dots, p_i, \dots, p_I\}$
  - $p(x = a_i) = p_i, p_i \geq 0$
  - $\sum_{a_i \in \mathcal{A}_X} P(x = a_i) = 1$
- Probability of a **subset**: if T is a subset of $\mathcal{A}_X$ then: $P(T) = P(x \in T) = \sum_{a_i \in T} P(x = a_i)$

Recap on probability theory

Definitions:

- X is a **random variable**, takes values $x \in \mathcal{A}_X = \{a_1, a_2, \ldots, a_i, \ldots, a_I\}$ with probabilities $\mathcal{P}_X = \{p_1, p_2, \ldots, p_i, \ldots, p_I\}$
  - $p(x = a_i) = p_i, p_i \geq 0$
  - $\sum_{a_i \in \mathcal{A}_X} P(x = a_i) = 1$
- Probability of a **subset**: if T is a subset of $\mathcal{A}_X$ then: $P(T) = P(x \in T) = \sum_{a_i \in T} P(x = a_i)$
- if XY is an ordered pair of variables where then $P(x, y)$ is the **joint probability** of x and y

Recap on probability theory

Definitions:

- X is a **random variable**, takes values $x \in \mathcal{A}_X = \{a_1, a_2, \ldots, a_i, \ldots, a_I\}$ with probabilities $\mathcal{P}_X = \{p_1, p_2, \ldots, p_i, \ldots, p_I\}$
  - $p(x = a_i) = p_i, p_i \geq 0$
  - $\sum_{a_i \in \mathcal{A}_X} P(x = a_i) = 1$
- Probability of a **subset**: if T is a subset of $\mathcal{A}_X$ then: $P(T) = P(x \in T) = \sum_{a_i \in T} P(x = a_i)$
- if XY is an ordered pair of variables where then $P(x, y)$ is the **joint probability** of x and y
- **Marginal** probability: $P(x = a_i) = \sum_{y \in \mathcal{A}_y} P(x = a_i, y)$

Recap on probability theory

Definitions:

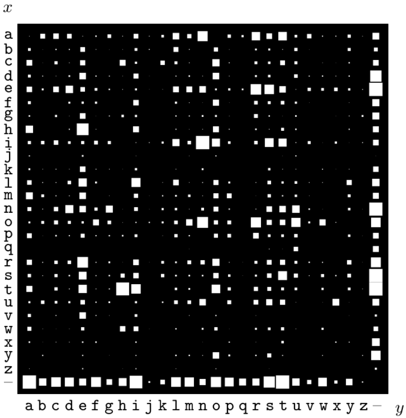
- X is a **random variable**, takes values $x \in \mathcal{A}_X = \{a_1, a_2, \ldots, a_i, \ldots, a_I\}$ with probabilities $\mathcal{P}_X = \{p_1, p_2, \ldots, p_i, \ldots, p_I\}$
  - $p(x = a_i) = p_i, p_i \geq 0$
  - $\sum_{a_i \in \mathcal{A}_X} P(x = a_i) = 1$
- Probability of a **subset**: if T is a subset of $\mathcal{A}_X$ then: $P(T) = P(x \in T) = \sum_{a_i \in T} P(x = a_i)$
- if XY is an ordered pair of variables where then $P(x, y)$ is the **joint probability** of x and y
- **Marginal** probability: $P(x = a_i) = \sum_{y \in \mathcal{A}_y} P(x = a_i, y)$
- **Conditional** probability:

$$P(x = a_i | y = b_j) = \frac{P(x = a_i, y = b_j)}{P(y = b_j)}, \text{if } P(y = b_j) \neq 0$$

Recap on probability theory

Example:

i	a_i	p_i
1	a	0.0575
2	b	0.0128
3	c	0.0263
4	d	0.0285
5	e	0.0913
6	f	0.0173
7	g	0.0133
8	h	0.0313
9	i	0.0599
10	j	0.0006
11	k	0.0084
12	l	0.0335
13	m	0.0235
14	n	0.0596
15	o	0.0689
16	p	0.0192
17	q	0.0008
18	r	0.0508
19	s	0.0567
20	t	0.0706
21	u	0.0334
22	v	0.0069
23	w	0.0119
24	x	0.0073
25	y	0.0164
26	z	0.0007
27	–	0.1928



Recap on probability theory

Rules of probability:

- Product rule $P(x, y) = P(x|y)P(y) = P(y|x)P(x)$

Recap on probability theory

Rules of probability:

- Product rule $P(x, y) = P(x|y)P(y) = P(y|x)P(x)$
- Sum rule $P(x) = \sum_y P(x, y) = \sum_y P(x|y)P(y)$

Recap on probability theory

Rules of probability:

- Product rule $P(x, y) = P(x|y)P(y) = P(y|x)P(x)$
- Sum rule $P(x) = \sum_y P(x, y) = \sum_y P(x|y)P(y)$
- Bayes theorem

$$P(y|x) = \frac{p(x|y)P(y)}{P(x)} = \frac{p(x|y)P(y)}{\sum_{y'} p(x|y')P(y')}$$

Recap on probability theory

Rules of probability:

- Product rule $P(x, y) = P(x|y)P(y) = P(y|x)P(x)$
- Sum rule $P(x) = \sum_y P(x, y) = \sum_y P(x|y)P(y)$
- Bayes theorem

$$P(y|x) = \frac{p(x|y)P(y)}{P(x)} = \frac{p(x|y)P(y)}{\sum_{y'} p(x|y')P(y')}$$

- Marginal independence: X and Y are independent $X \perp\!\!\!\perp Y|\emptyset$ if and only if

$$P(x, y) = P(x)P(y)$$

Recap on probability theory

Rules of probability:

- Product rule $P(x, y) = P(x|y)P(y) = P(y|x)P(x)$
- Sum rule $P(x) = \sum_y P(x, y) = \sum_y P(x|y)P(y)$
- Bayes theorem

$$P(y|x) = \frac{p(x|y)P(y)}{P(x)} = \frac{p(x|y)P(y)}{\sum_{y'} p(x|y')P(y')}$$

- Marginal independence: X and Y are independent $X \perp\!\!\!\perp Y|\emptyset$ if and only if

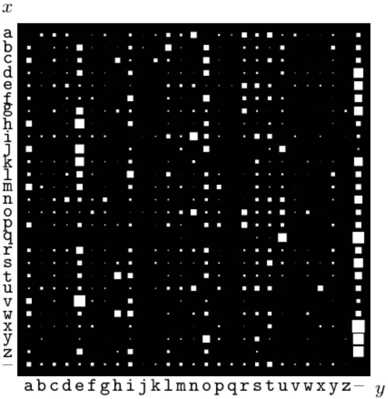
$$P(x, y) = P(x)P(y)$$

- Conditional independence: X and Y are independent given Z $X \perp\!\!\!\perp Y|Z$ if and only if

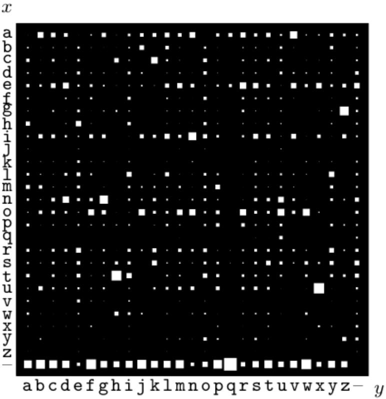
$$P(x, y|z) = P(x|z)P(y|z)$$

Recap on probability theory

Example:



(a) $P(y \mid x)$



(b) $P(x \mid y)$

Are x and y independent?

Recap on probability theory

Example Bayes (I/II):

Example 2.3. Jo has a test for a nasty disease. We denote Jo's state of health by the variable a and the test result by b.

$$\begin{array}{ll} a = 1 & \text{Jo has the disease} \\ a = 0 & \text{Jo does not have the disease.} \end{array} \tag{2.12}$$

The result of the test is either 'positive' ($b = 1$) or 'negative' ($b = 0$); the test is 95% reliable: in 95% of cases of people who really have the disease, a positive result is returned, and in 95% of cases of people who do not have the disease, a negative result is obtained. The final piece of background information is that 1% of people of Jo's age and background have the disease.

OK – Jo has the test, and the result is positive. What is the probability that Jo has the disease?

Recap on probability theory

Example Vaccinations:
A negationist tells you that 60% of the people in the hospital with COVID are vaccinated and 40% are not. Therefore you should not vaccinate. What is wrong with his argumentation?

Recap on probability theory

Example Vaccinations:
A negationist tells you that 60% of the people in the hospital with COVID are vaccinated and 40% are not. Therefore you should not vaccinate. What is wrong with his argumentation?

1. $P(v = 1|h = 1)$ is just one of the three pieces of information. We also need to consider the probability of vaccinated and the probability of being in the hospital.

$$p(h = 1|v = 1) = \frac{p(v = 1|h = 1)p(h = 1)}{p(v = 1)}$$

Recap on probability theory

Example Vaccinations:

A negationist tells you that 60% of the people in the hospital with COVID are vaccinated and 40% are not. Therefore you should not vaccinate. What is wrong with his argumentation?

1. $P(v = 1|h = 1)$ is just one of the three pieces of information. We also need to consider the probability of vaccinated and the probability of being in the hospital.

$$p(h = 1|v = 1) = \frac{0.6 \cdot 0.0001}{0.8} = 7.5 \cdot 10^{-5}$$

Recap on probability theory

Example Vaccinations:
A negationist tells you that 60% of the people in the hospital with COVID are vaccinated and 40% are not. Therefore you should not vaccinate. What is wrong with his argumentation?

1. $P(v = 1|h = 1)$ is just one of the three pieces of information. We also need to consider the probability of vaccinated and the probability of being in the hospital.

$$p(h = 1|v = 0) = \frac{0.4 \cdot 0.0001}{0.2} = 2 \cdot 10^{-4}$$

Inverse Probabilities

Forward and inverse probabilities

- Forward probabilities: given a *generative model*, compute distribution of *data produced* by the model

Example:

Exercise 2.4.[2, p.40] An urn contains K balls, of which B are black and $W = K - B$ are white. Fred draws a ball at random from the urn and replaces it, N times.

(a) What is the probability distribution of the number of times a black ball is drawn, n_B?

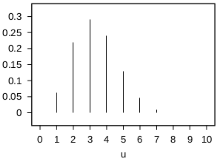
Forward and inverse probabilities

- Inverse probabilities: given a *generative model*, compute distribution of *hidden variables* in the model, from the observed data

Example 2.6. There are eleven urns labelled by $u \in \{0, 1, 2, \ldots, 10\}$, each containing ten balls. Urn u contains u black balls and $10 - u$ white balls. Fred selects an urn u at random and draws N times with replacement from that urn, obtaining n_B blacks and $N - n_B$ whites. Fred's friend, Bill, looks on. If after $N = 10$ draws $n_B = 3$ blacks have been drawn, what is the probability that the urn Fred is using is urn u, from Bill's point of view? (Bill doesn't know the value of u.)

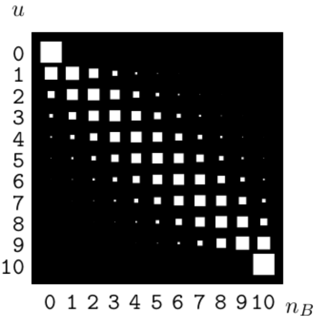
Forward and inverse probabilities

Forward and inverse probabilities



u	$P(u \mid n_B = 3, N)$
0	0
1	0.063
2	0.22
3	0.29
4	0.24
5	0.13
6	0.047
7	0.0099
8	0.00086
9	0.0000096
10	0

Figure 2.6. Conditional probability of u given $n_B = 3$ and $N = 10$.



Forward and inverse probabilities

Terminology

- $P(u)$ *prior over* u (encodes prior knowledge about our model)

Forward and inverse probabilities

Terminology

- $P(u)$ *prior over u* (encodes prior knowledge about our model)
- $P(n_B|u, N)$ *likelihood of u* (NOT likelihood of n_B!!)

Forward and inverse probabilities

Terminology

- $P(u)$ *prior over* u (encodes prior knowledge about our model)
- $P(n_B|u, N)$ *likelihood of* u (NOT likelihood of n_B!!)
- $P(u|n_B, N)$ *posterior probability* of u given n_B

Forward and inverse probabilities

Terminology

- $P(u)$ *prior over* u (encodes prior knowledge about our model)
- $P(n_B|u, N)$ *likelihood of* u (NOT likelihood of n_B!!)
- $P(u|n_B, N)$ *posterior probability* of u given n_B
- $P(n_B|N)$ *evidence or marginal likelihood* of n_B

Forward and inverse probabilities

Terminology

- $P(u)$ *prior over* u (encodes prior knowledge about our model)
- $P(n_B|u, N)$ *likelihood of* u (NOT likelihood of n_B!!)
- $P(u|n_B, N)$ *posterior probability* of u given n_B
- $P(n_B|N)$ *evidence or marginal likelihood* of n_B

Bayes I

$$P(\boldsymbol{\theta}|\mathcal{D},\mathcal{H}) = \frac{P(\mathcal{D}|\boldsymbol{\theta},\mathcal{H})P(\boldsymbol{\theta}|\mathcal{H})}{P(\mathcal{D}|\mathcal{H})}$$

Forward and inverse probabilities

Inverse probability and prediction

- Involves *marginalizing* over possible values of the hypothesis

Example 2.6 (continued). Assuming again that Bill has observed $n_B = 3$ blacks in $N = 10$ draws, let Fred draw another ball from the same urn. What is the probability that the next drawn ball is a black? [You should make use of the posterior probabilities in figure 2.6.]

Forward and inverse probabilities

1st set of exercises for next week. From chap. 2 of D. Mackay

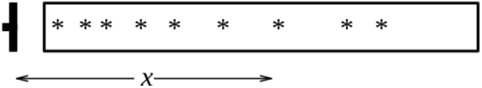
- 2.7.
- 2.8.
- 2.10.
- 2.11.

Learning as Inference (Bayes I)

Learning as inference

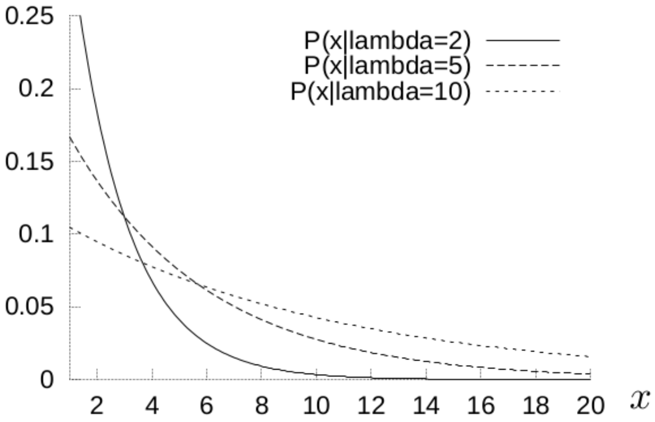
Exercise 3.3.[3, p.48] Inferring a decay constant

Unstable particles are emitted from a source and decay at a distance x, a real number that has an exponential probability distribution with characteristic length λ. Decay events can be observed only if they occur in a window extending from $x = 1\,\text{cm}$ to $x = 20\,\text{cm}$. N decays are observed at locations $\{x_1, \ldots, x_N\}$. What is λ?



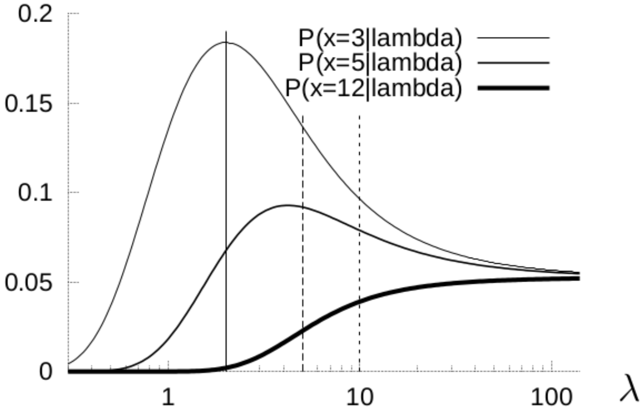
Learning as inference

Learning as inference



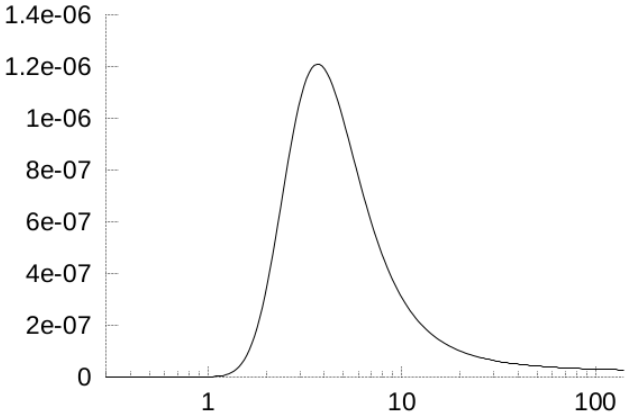
The probability density $p(x|\lambda)$ (a truncated exponential)

Learning as inference



The likelihood function $p(x|\lambda)$ for one data point

Learning as inference



The likelihood function for $\mathcal{D} = \{1.5, 2, 3, 4, 5, 12\}$

Learning as inference

Bayesian Linear Regression

- Remember linear regression: find w that minimizes

$$L_S(w) = \frac{1}{m} \sum_{i=1}^{m} (y_i - w^\top x_i)^2$$

Learning as inference

Bayesian Linear Regression

- Data points generated as noisy targets $y_i \sim w^\top x_i + \eta$
- If noise is Gaussian, $\eta \sim \mathcal{N}(0, \sigma^2)$, the model generates y_i

$$
\begin{aligned}
p(y_i|x_i, w) &= \mathcal{N}(w^\top x_i, \sigma^2) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - w^\top x_i)^2\right)
\end{aligned}
$$

- For m i.i.d. datapoints, the likelihood becomes

$$
p(\mathcal{D}|w) = \prod_{i=1}^{m} p(y_i|x_i, w) p(x_i)
$$

Learning as inference

Bayesian Linear Regression

- Ignoring $1/\sigma^2$ and the input distribution $p(x_i)$, taking log

$$\log p(\mathcal{D}|w) = -\sum_{i=1}^{m} (y_i - w^\top x_i)^2$$

Learning as inference

Bayesian Linear Regression

- Ignoring $1/\sigma^2$ and the input distribution $p(x_i)$, taking log

$$\log p(\mathcal{D}|w) = -\sum_{i=1}^{m}(y_i - w^\top x_i)^2$$

- **Minimizing squared error is equivalent to maximizing the likelihood under Gaussian noisy outputs**

Learning as inference

Bayesian Linear Regression

- Ignoring $1/\sigma^2$ and the input distribution $p(x_i)$, taking log

$$\log p(\mathcal{D}|w) = -\sum_{i=1}^{m}(y_i - w^\top x_i)^2$$

- **Minimizing squared error is equivalent to maximizing the likelihood under Gaussian noisy outputs**
- What are we missing?

Learning as inference

Bayesian Linear Regression

- The priors!

Learning as inference

Bayesian Linear Regression

- The priors!
- For Gaussian priors $p(w|\lambda) \sim \mathcal{N}(0, 1/\lambda^2)$, posterior is

$$\log p(w|\mathcal{D}, \lambda) = -\sum_{i=1}^{m}(y_i - w^\top x_i)^2 - \lambda w^\top w + \text{const}$$

Learning as inference

Bayesian Linear Regression

- The priors!
- For Gaussian priors $p(w|\lambda) \sim \mathcal{N}(0, 1/\lambda^2)$, posterior is

$$\log p(w|\mathcal{D}, \lambda) = -\sum_{i=1}^{m}(y_i - w^\top x_i)^2 - \lambda w^\top w + \text{const}$$

- Remember linear regression with regularization

$$L_{aug}(w) = \frac{1}{m}\sum_{i=1}^{m}(y_i - w^\top x_i)^2 + \frac{\lambda}{m} w^\top w$$

Learning as inference

Bayesian Linear Regression

- The priors!
- For Gaussian priors $p(w|\lambda) \sim \mathcal{N}(0, 1/\lambda^2)$, posterior is

$$\log p(w|\mathcal{D}, \lambda) = -\sum_{i=1}^{m}(y_i - w^\top x_i)^2 - \lambda w^\top w + \text{const}$$

- Remember linear regression with regularization

$$L_{aug}(w) = \frac{1}{m}\sum_{i=1}^{m}(y_i - w^\top x_i)^2 + \frac{\lambda}{m} w^\top w$$

- **The prior plays the role of regularization**

Model Comparison as Inference (Bayes II)

Model Comparison

How to choose between models / hypothesis?

Bayes II

$$P(\mathcal{H}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{H})P(\mathcal{H})}{P(\mathcal{D})}$$

Model Comparison

Bayes I

$$P(\boldsymbol{\theta}|\mathcal{D},\mathcal{H}) = \frac{P(\mathcal{D}|\boldsymbol{\theta},\mathcal{H})P(\boldsymbol{\theta}|\mathcal{H})}{P(\mathcal{D}|\mathcal{H})}$$

Bayes II

$$P(\mathcal{H}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{H})P(\mathcal{H})}{P(\mathcal{D})}$$

Note that $P(\mathcal{D}|\mathcal{H})$ corresponds to the evidence of Bayes I (a.k.a. *marginal likelihood*)

Model Comparison

3.2 The bent coin

A bent coin is tossed F times; we observe a sequence $\mathbf{s}$ of heads and tails (which we'll denote by the symbols a and b). We wish to know the bias of the coin, and predict the probability that the next toss will result in a head.

- Hypothesis $\mathcal{H}_1$ assumes that there is an unknown bias (parameter) we want to infer

Model Comparison

Model Comparison

- Posterior is

$$p(p_a|\mathbf{s}, \mathcal{H}_1) = \frac{p_a^{F_a}(1-p_a)^{F_b}}{p(\mathbf{s}|F, \mathcal{H}_1)}$$

- Evidence is

$$p(\mathbf{s}|\mathcal{H}_1) = \int_0^1 dp_a p_a^{F_a}(1-p_a)^{F_b} = \frac{F_a!F_b!}{(F_a+F_b+1)!}$$

Model Comparison

- Predictions

$$
\begin{aligned}
p(\mathrm{a}|\mathbf{s}, \mathcal{H}_1) &= \int_0^1 dp_a p(\mathrm{a}|p_a, \mathcal{H}_1) p(p_a|\mathbf{s}, \mathcal{H}_1) \\
&= \frac{F_a + 1}{F_a + F_b + 2}
\end{aligned}
$$

Model Comparison

3.3 The bent coin and model comparison

Imagine that a scientist introduces another theory for our data. He asserts that the source is not really a bent coin but is really a perfectly formed die with one face painted heads ('a') and the other five painted tails ('b'). Thus the parameter $p_{\mathtt{a}}$, which in the original model, $\mathcal{H}_1$, could take any value between 0 and 1, is according to the new hypothesis, $\mathcal{H}_0$, not a free parameter at all; rather, it is equal to 1/6. [This hypothesis is termed $\mathcal{H}_0$ so that the suffix of each model indicates its number of free parameters.]

Model Comparison

- The likelihood under the (simpler) hypothesis $\mathcal{H}_0$ is

$$p(\mathbf{s}|\mathcal{H}_0) = (1/6)^{F_a}(1 - 1/6)^{F_b}$$

- The ratio of posteriors is

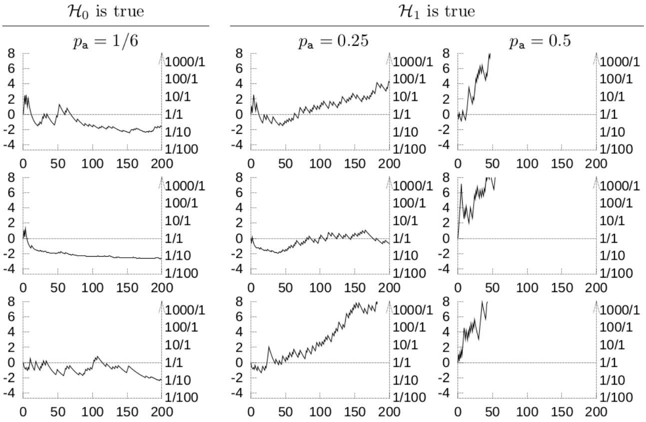
$$\frac{p(\mathcal{H}_1|\mathbf{s})}{p(\mathcal{H}_0|\mathbf{s})} = \frac{F_a!F_b!}{(F_a + F_b + 1)!}\frac{1}{(1/6)^{F_a}(1 - 1/6)^{F_b}}$$

Model Comparison

F	Data (F_a, F_b)	$\frac{P(\mathcal{H}_1 \mid \mathbf{s}, F)}{P(\mathcal{H}_0 \mid \mathbf{s}, F)}$	
6	$(5, 1)$	222.2	
6	$(3, 3)$	2.67	
6	$(2, 4)$	0.71	$= 1/1.4$
6	$(1, 5)$	0.356	$= 1/2.8$
6	$(0, 6)$	0.427	$= 1/2.3$
20	$(10, 10)$	96.5	
20	$(3, 17)$	0.2	$= 1/5$
20	$(0, 20)$	1.83	

Some values of $\frac{p(\mathcal{H}_1 | \mathbf{s})}{p(\mathcal{H}_0 | \mathbf{s})}$ for different data

Model Comparison



Behavior as a function of the size of the data

Model Comparison

Example: three doors

On a game show, a contestant is told the rules as follows:

> There are three doors, labelled 1, 2, 3. A single prize has been hidden behind one of them. You get to select one door. Initially your chosen door will *not* be opened. Instead, the gameshow host will open one of the other two doors, and *he will do so in such a way as not to reveal the prize.* For example, if you first choose door 1, he will then open one of doors 2 and 3, and it is guaranteed that he will choose which one to open so that the prize will not be revealed.
>
> At this point, you will be given a fresh choice of door: you can either stick with your first choice, or you can switch to the other closed door. All the doors will then be opened and you will receive whatever is behind your final choice of door.

Imagine that the contestant chooses door 1 first; then the gameshow host opens door 3, revealing nothing behind the door, as promised. Should the contestant (a) stick with door 1, or (b) switch to door 2, or (c) does it make no difference?

Model Comparison

Example: three doors

- $\mathcal{H}_i$: car is at door i. $i \in \{1, 2, 3\}$.
- Likelihoods

$$\left| \begin{array}{c} P(D=2 \mid \mathcal{H}_1) = 1/2 \\ P(D=3 \mid \mathcal{H}_1) = 1/2 \end{array} \right| \begin{array}{c} P(D=2 \mid \mathcal{H}_2) = 0 \\ P(D=3 \mid \mathcal{H}_2) = 1 \end{array} \left| \begin{array}{c} P(D=2 \mid \mathcal{H}_3) = 1 \\ P(D=3 \mid \mathcal{H}_3) = 0 \end{array} \right|$$

- Posterior

$$P(\mathcal{H}_i \mid D=3) = \frac{P(D=3 \mid \mathcal{H}_i) P(\mathcal{H}_i)}{P(D=3)}$$

$$\left| \; P(\mathcal{H}_1 \mid D=3) \;=\; 1/3 \;\right| \; P(\mathcal{H}_2 \mid D=3) \;=\; 2/3 \;\left| \; P(\mathcal{H}_3 \mid D=3) \;=\; 0. \;\right|$$

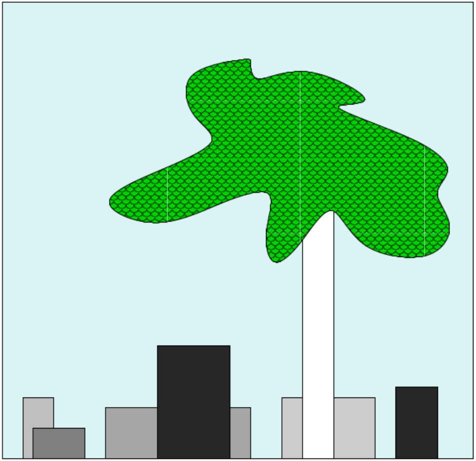
Model Comparison

Model Comparison

2nd set of exercises for next week. From chap. 3 of D. Mackay

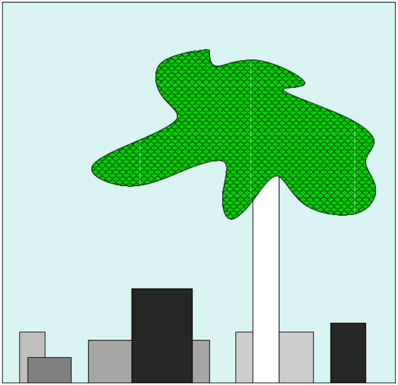
- 3.5.
- 3.10.
- 3.12.
- 3.14.

Occam's razor

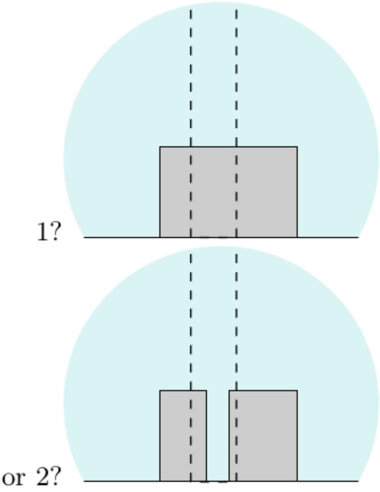


How many boxes are in the picture?

Occam's razor

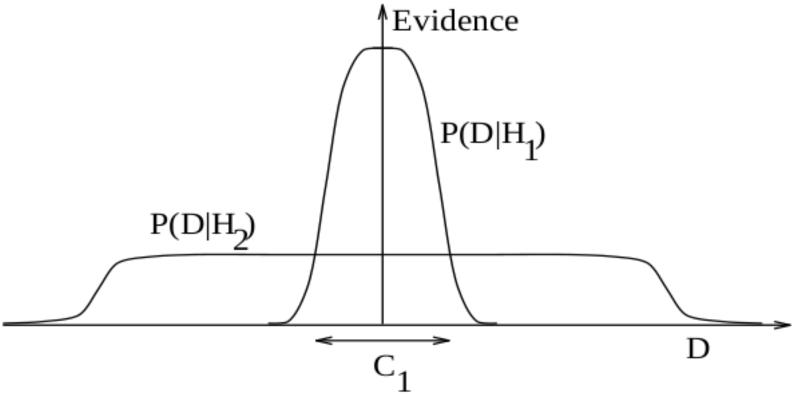


How many boxes are in the picture?



Occam's razor

- Accept the *simplest* explanation that fits the data
- Bayesian inference embodies Occam's razor *automatically*



Occam's razor

Example: sequence of numbers

- Given the sequence:

$$-1, 3, 7, 11$$

Occam's razor

Example: sequence of numbers

- Given the sequence:

$$-1, 3, 7, 11$$

- What are the next two numbers? What is the generating process?

Occam's razor

Example: sequence of numbers

- Given the sequence:

$$-1, 3, 7, 11$$

- What are the next two numbers? What is the generating process?
- Option 1: $(15, 19, \ldots)$ *start from* -1*, and add* 1 *to the previous number*

Occam's razor

Example: sequence of numbers

- Given the sequence:

$$-1, 3, 7, 11$$

- What are the next two numbers? What is the generating process?
- Option 1: $(15, 19, \ldots)$ *start from* -1*, and add* 1 *to the previous number*
- Option 2: $(-19.9, 1043.8, \ldots)$ *start from* -1*, use the previous number* x *to get the new one according to*

$$-x^3/11 + 9/11x^2 + 23/11$$

Occam's razor

$\mathcal{H}_a$ – the sequence is an *arithmetic* progression, 'add n', where n is an integer.

$\mathcal{H}_c$ – the sequence is generated by a *cubic* function of the form $x \rightarrow cx^3 + dx^2 + e$, where c, d and e are fractions.

Occam's razor

$\mathcal{H}_a$ – the sequence is an *arithmetic* progression, 'add n', where n is an integer.

$\mathcal{H}_c$ – the sequence is generated by a *cubic* function of the form $x \to cx^3 + dx^2 + e$, where c, d and e are fractions.

- Model $\mathcal{H}_a$ has **two** parameters: first number, and n

$$P(D \mid \mathcal{H}_a) = \frac{1}{101}\frac{1}{101} = 0.00010.$$

- Model $\mathcal{H}_c$ has **four** parameters: first number, c, d, and e

$$\begin{aligned} P(D \mid \mathcal{H}_c) &= \left(\frac{1}{101}\right)\left(\frac{4}{101}\frac{1}{50}\right)\left(\frac{4}{101}\frac{1}{50}\right)\left(\frac{2}{101}\frac{1}{50}\right) \\ &= 0.0000000000025 = 2.5 \times 10^{-12}. \end{aligned}$$