Tutorial 1 (CS 412)

Note: Question [x.y] refers to question y of the exercises in Chapter x.

• Probability basics

Question 0. TRUE or FALSE

X independent of Y and Y independent of Z implies that X is independent of Z.

FALSE (say, X = Z)

Question 1. Consider two medical tests, A and B, for a virus. Test A is 95% effective at recognizing the virus when it is present, but has a 10% false positive rate (indicating that the virus is present, when it is not). Test B is 90% effective at recognizing the virus, but has a 5% false positive rate. The two tests use independent methods of identifying the virus. The virus is carried by 1% of all people. Say that a person is tested for the virus using only one of the tests, and that test comes back positive for carrying the virus. Which test returning positive is more indicative of someone really carrying the virus? Justify your answer mathematically.

Let V be the statement that the patient has the virus, and A and B the statements that the medical tests A and B returned positive, respectively. The problem statement gives:

$$\begin{aligned} P(V) &= 0.01 \\ P(A|V) &= 0.95 \\ P(A|\neg V) &= 0.10 \\ P(B|V) &= 0.90 \\ P(B|\neg V) &= 0.05 \end{aligned}$$

The test whose positive result is more indicative of the virus being present is the one whose posterior probability, $P(V|A)$ or $P(V|B)$ is largest. One can compute these probabilities directly from the information given, finding that $P(V|A) = 0.0876$ and $P(V|B) = 0.1538$, so B is more indicative.

Equivalently, the questions is asking which test has the highest posterior odds ratio $P(V|A)/P(\neg V|A)$. From the odd form of Bayes theorem:

$$\frac{P(V|A)}{P(\neg V|A)} = \frac{P(A|V)}{P(A|\neg V)}\frac{P(V)}{P(\neg V)}$$

we see that the ordering is independent of the probability of V, and that we just need to compare the likelihood ratios $P(A|V)/P(A|\neg V) = 9.5$ and $P(B|V)/P(V|\neg V) = 18$ to find the answer.

Question 2. Let X, Y, Z be Boolean random variables. Label the eight entries in the joint distribution P(X, Y, Z) as *a* through *h*. Express the statement that X and Y are conditionally independent given Z, as a set of equations relating *a* through *h*. How many non-redundant equations are there?

Solution: Let the probabilities be as follows:

x	y	z	$P(x,y,z)$
F	F	F	a
F	F	T	b
F	T	F	c
F	T	T	d
T	F	F	e
T	F	T	f
T	T	F	g
T	T	T	h

Conditional independence asserts that

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

which we can rewrite in terms of the joint distribution using the definition of conditional probability and marginals:

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

$$\mathbf{P}(X,Y,Z) = \frac{\mathbf{P}(X,Z)\mathbf{P}(Y,Z)}{\mathbf{P}(Z)} = \frac{\left(\sum_y \mathbf{P}(X,y,Z)\right)\left(\sum_x \mathbf{P}(x,Y,Z)\right)}{\sum_{x,y} \mathbf{P}(x,y,Z)} .$$

Now we instantiate X, Y, Z in all 8 ways to obtain the following 8 equations:

$$\begin{aligned}
a &= (a+c)(a+e)/(a+c+e+g) \text{ or } ag = ce \\
b &= (b+d)(b+f)/(b+d+f+h) \text{ or } bh = df \\
c &= (a+c)(c+g)/(a+c+e+g) \text{ or } ce = ag \\
d &= (b+d)(d+h)/(b+d+f+h) \text{ or } df = bh \\
e &= (e+g)(a+e)/(a+c+e+g) \text{ or } ce = ag \\
f &= (f+h)(b+f)/(b+d+f+h) \text{ or } df = bh \\
g &= (e+g)(c+g)/(a+c+e+g) \text{ or } ag = ce \\
h &= (f+h)(d+h)/(b+d+f+h) \text{ or } bh = df .
\end{aligned}$$

Thus, there are only 2 nonredundant equations, $ag = ce$ and $bh = df$. This is what we would expect: the general distribution requires $8 - 1 = 7$ parameters, whereas the Bayes net with Z as root and X and Y as conditionally indepednent children requires 1 parameter for Z and 2 each for X and Y, or 5 in all. Hence the conditional independence assertion removes two degrees of freedom.

Question 3. Show that the three forms of independence are equivalent:
P(a | b) = P(a) or P(b | a) = P(b) or P(a ∧ b) = P(a) P(b) .

Solution:

Independence is symmetric (that is, *a* and *b* are independent if and only if *b* and *a* are independent) so $P(a|b) = P(a)$ is the same as $P(b|a) = P(b)$. So we need only prove that $P(a|b) = P(a)$ is equivalent to $P(a \wedge b) = P(a)P(b)$. The product rule, $P(a \wedge b) = P(a|b)P(b)$, can be used to rewrite $P(a \wedge b) = P(a)P(b)$ as $P(a|b)P(b) = P(a)P(b)$, which simplifies to $P(a|b) = P(a)$.

Question 4.

a) Give an example where random variables X and Y are independent, but they are no longer independent given Z. You need to write out the joint distribution in detail.

Solution (not unique):

x	y	z	P(x, y, z)
0	0	0	0.25
0	1	1	0.25
1	0	1	0.25
1	1	0	0.25

and all other combinations of (x, y, z) have probability 0. Clearly X and Y are independent after Z is marginalized out (all combinations of X and Y have probability 0.25).

Now given z=1, we see x and y has to be different, and given z=0, x and y has to be same. Hence x and y are not independent given z.

b) Give an example where random variables X and Y are not independent, but they become independent given Z. You need to write out the joint distribution in detail.

Solution (not unique):

Suppose there are two coins: a regular coin and a two-headed coin $(P(H) = 1)$ one choose a coin uniformly at random and toss it twice.

X = Coin 1 (regular) has been selected.

Y = first flip is H

Z = second flip is H

If we know Y has occurred, we would guess that it is more likely that we have chosen Coin 2 than Coin 1. This in turn increases the conditional probability that Z occurs. This suggests that Y and Z are not independent. On the other hand, $Y \perp Z|X$. The joint probability table is

X	Y	Z	probability
T	T	T	$0.5 * 0.5 * 0.5 = 0.125$
T	T	F	$0.5 * 0.5 * 0.5 = 0.125$
T	F	T	$0.5 * 0.5 * 0.5 = 0.125$
T	F	F	$0.5 * 0.5 * 0.5 = 0.125$
F	T	T	$0.5 * 1 * 1 = 0.5$
F	T	F	$0.5 * 1 * 0 = 0$
F	F	T	0
F	F	F	0

(1)

P(Y=T) = 0.75

P(Z=F) = 0.25

P(Y=T, Z=F) = 0.125 which is different from

P(Y=T) * P(Z=T). So Y and Z are not independent.

(2)

Verify: P(Y,Z|X) = P(Y|X) P(Z|X)

for all assignments of X, Y, and Z.

Question 5.

Consider the joint distribution, $P(x, y, z)$ where each is a binary-valued variable. If $X \perp Y$, how many free parameters does this distribution have? (Hint: without this independence property there are $2^3 - 1$ free parameters.)

Solution:

Similar to Question 2, if X is independent on Y, we only get one nonredundant equation. So we have 6 free parameters.

To see it more concretely, note P(X, Y, Z) = P(Z|X,Y)P(X)P(Y). So four numbers for P(Z|X,Y), one number for P(X) and one number for P(Y).

Question 6. Multiple choice

a) Consider a continuous-valued random variable, X, that can take values from the set $\mathcal{X}$. Which of the following is always true (i.e., for any choice of set $\mathcal{X}$ and distributions over the random variable X) for probability mass functions P and probability density functions f:

(a) $P(x) > 0$ for some $x \in \mathcal{X}$
(b) $f(x) > 0$ for some $x \in \mathcal{X}$
(c) $P(X \in \mathcal{X}') \leq 1$ for all $\mathcal{X}' \subseteq \mathcal{X}$
(d) $f(x) \leq 1$ for all $x \in \mathcal{X}$

Solution: (c).

Here P is probability, f is the density function.

(a) For a single point x in continuous valued distribution, P(x) is always 0.

(b) For some x, its probability density function can be 0.

(d) density can be well above 1; think of f(0) for a Gaussian distribution with zero mean and variance 10^{-6}.

b) Assume you have a fair coin and flip it three times, X_1, X_2, X_3 are three random variable denote the result of each flip separately ($X_i = 1$ if the result of flip is head, $X_i = 0$ otherwise) which of the following is/are correct?

(a) $P(\min(X_1, X_2, X_3) = 0) < P(\max(X_1, X_2, X_3) = 1)$
(b) $P(X_1 < X_2 \leq X_3) = P(X_1 > X_2 \geq X_3)$
(c) $P(\max(X_1, X_2, X_3) > \min(X_1, X_2, X_3)) = 1$
(d) $P(X_1^2 + X_2^2 + X_3^2 = X_1 + X_2 + X_3) = 1$

Solution: (b)(d). List all cases noting that because the coin is fair all combinations of the values of X_1, X_2, X_3 have probability 1/8.

(a) They should have the same probability.

(c) There is still 1/4 chance that all these variables are equal (all 0 or all 1).

Question 7. Suppose you are given a bag containing *n* unbiased coins. You are told that *n* – 1 of these coins are normal, with heads on one side and tails on the other, whereas one coin is a fake, with heads on both sides.

a. Suppose you reach into the bag, pick out a coin at random, flip it, and get a head. What is the (conditional) probability that the coin you chose is the fake coin?

b. Suppose you continue flipping the coin for a total of *k* times after picking it and see *k* heads. Now what is the conditional probability that you picked the fake coin?

c. Suppose you wanted to decide whether the chosen coin was fake by flipping it *k* times. The decision procedure returns *fake* if all *k* flips come up heads; otherwise it returns *normal*. What is the (unconditional) probability that this procedure makes an error?

a. A typical "counting" argument goes like this: There are n ways to pick a coin, and 2 outcomes for each flip (although with the fake coin, the results of the flip are indistinguishable), so there are $2n$ total atomic events, each equally likely. Of those, only 2 pick the fake coin, and $2 + (n - 1)$ result in heads. So the probability of a fake coin given heads, $P(fake|heads)$, is $2/(2 + n - 1) = 2/(n + 1)$.

Often such counting arguments go astray when the situation gets complex. It may be better to do it more formally:

$$\begin{aligned}\mathbf{P}(Fake|heads) &= \alpha\mathbf{P}(heads|Fake)\mathbf{P}(Fake)\\ &= \alpha\langle 1.0, 0.5\rangle\langle 1/n, (n-1)/n\rangle\\ &= \alpha\langle 1/n, (n-1)/2n\rangle\\ &= \langle 2/(n+1), (n-1)/(n+1)\rangle\end{aligned}$$

b. Now there are $2^k n$ atomic events, of which 2^k pick the fake coin, and $2^k + (n-1)$ result in heads. So the probability of a fake coin given a run of k heads, $P(fake|heads^k)$, is $2^k/(2^k + (n-1))$. Note this approaches 1 as k increases, as expected. If $k = n = 12$, for example, than $P(fake|heads^{10}) = 0.9973$.

Doing it the formal way:

$$\begin{aligned}\mathbf{P}(Fake|heads^k) &= \alpha\mathbf{P}(heads^k|Fake)\mathbf{P}(Fake)\\ &= \alpha\langle 1.0, 0.5^k\rangle\langle 1/n, (n-1)/n\rangle\\ &= \alpha\langle 1/n, (n-1)/2^k n\rangle\\ &= \langle 2^k/(2^k+n-1), (n-1)/(2^k+n-1)\rangle\end{aligned}$$

c. The procedure makes an error if and only if a fair coin is chosen and turns up heads k times in a row. The probability of this

$$P(heads^k|\neg fake)P(\neg fake) = (n-1)/2^k n\,.$$

Question 8. In the Chicago weather examples from the lecture, suppose

P(Cold) = 0.7	P(Snow | Cold, Cloudy) = 0.8	P(Snow | Cold, Clear) = 0
P(Cloudy) = 0.6	P(Snow | Warm, Cloudy) = 0	P(Snow | Warm, Clear) = 0

Suppose Temperature is independent of Sky. Then compute P(cold, cloudy | not snow).

Solution:

$$P(\text{cold, cloudy} \mid \text{not snow}) = \frac{P(\text{cold,cloudy,not snow})}{P(\text{not snow})}$$

The numerator is equal to $P(\text{not snow} \mid \text{cold, cloudy}) \times P(\text{cold, cloudy})$. Noting that $P(\text{cold, cloudy}) = P(\text{cold}) \times P(\text{cloudy})$ by independence, we find that the numerator is equal to 0.2*0.7*0.6 = 0.084.

The denominator is equal to

$$P(\text{cold, cloudy, not snow}) + P(\text{warm, cloudy, not snow})$$

$$+ P(\text{cold, clear, not snow}) + P(\text{warm, clear, not snow})$$

= P(not snow | cold, cloudy) × P(cold, cloudy) + P(not snow | warm, cloudy) × P(warm, cloudy)

+ P(not snow | cold, clear) × P(cold, clear) + P(not snow | warm, clear) × P(warm, clear)

= 0.2*0.7*0.6 + 1*0.3*0.6 + 1*0.7*0.4 + 1*0.3*0.4

= 0.664.

So P(cold, cloudy | not snow) = 0.084 / 0.664 = 0.127.

- **Supervised learning**

Questions [2.1-2.5]. Solutions are available in the textbook.

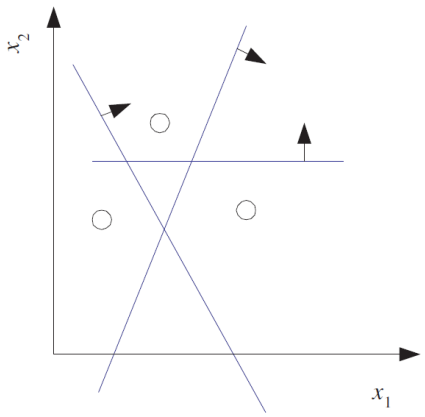
Question [2.7] (you need to understand it, but deriving it by yourself is optional)

We take the derivative of the sum of squared errors with respect to the two parameters, set them equal to 0, and solve these two equations in two unknowns:

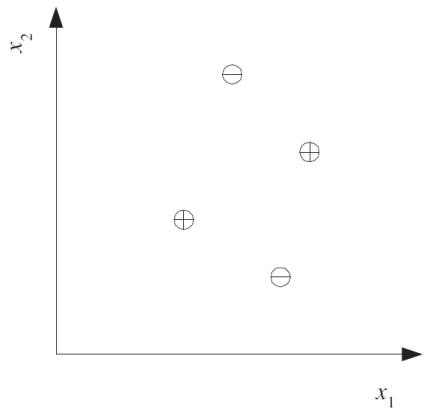
$$
\begin{aligned}
E(w_1, w_0|\mathcal{X}) &= \frac{1}{N}\sum_{i=1}^{N}\left[r^t - (w_1 x^t + w_0)\right]^2 \\
\frac{\partial E}{\partial w_0} &= \sum_t \left[r^t - (w_1 x^t + w_0)\right] = 0 \\
\sum_t r^t &= w_1 \sum_t x^t + N w_0 \\
w_0 &= \sum_t r^t/N - w_1 \sum_t x^t/N = \overline{r} - w_1 \overline{x} \\
\frac{\partial E}{\partial w_1} &= \sum_t \left[r^t - (w_1 x^t + w_0)\right] x^t = 0 \\
\sum_t r^t x^t &= w_1 \sum_t (x^t)^2 + w_0 \sum_t x^t \\
\sum_t r^t x^t &= w_1 \sum_t (x^t)^2 + (\overline{r} - w_1 \overline{x}) \sum_t x^t \\
\sum_t r^t x^t &= w_1 \left(\sum_t (x^t)^2 - \overline{x}\sum_t x^t\right) + \overline{r}\sum_t x^t \\
\sum_t r^t x^t &= w_1 \left(\sum_t (x^t)^2 - \overline{x} N \overline{x}\right) + \overline{r} N \overline{x} \\
w_1 &= \frac{\sum_t r^t x^t - \overline{x}\overline{r}N}{\sum_t (x^t)^2 - N\overline{x}^2}
\end{aligned}
$$

Question [2.8]

As we see below for all possible labeling of three points, there exist a line to separate positive and negative examples. With four points, no matter how we place these four points in two dimensions, there is at least one labeling where we cannot draw a line such that on one side lie all the positives and on the other lie all the negatives.



All possible labelings of three points can be separated using a line.



These four points cannot be separated using a line.

Question [2.11]

If we have an instance that is surrounded by many instances with a different label, then it is most probably mislabeled. By using a neighbor-based method (we will see in chapter 8), we can check for this.

- **Bayesian decision theory**

Question [3.1 to 3.4]. Solutions are available in the textbook.

- **Experiment Design and Cross validation**

Q [19.1] See textbook

Question 1. Suppose you are running a learning experiment on a new algorithm for Boolean classification. You have a data set consisting of 100 positive and 100 negative examples. You plan to use leave-one-out cross-validation and compare your algorithm to a baseline function, a simple majority classifier. (A majority classifier is given a set of training data and then always outputs the class that is in the majority in the training set, regardless of the input.) You expect the majority classifier to score about 50% on leave-one-out cross-validation, but to your surprise, it scores zero every time. Can you explain why?

Solution: If we leave out an example of one class, then the majority of the remaining examples are of the other class, so the majority classifier will always predict the wrong answer.

Question 2. To evaluate the usefulness of a classifier, the best metric to use is (circle one):
(a) Classification accuracy on the training dataset
(b) Classification accuracy on a withheld dataset not used for training
(c) Classification accuracy on a dataset created by combined the training dataset with all additional withheld data
(d) Classification accuracy on a random subset of the training dataset

Solution: (b)

Question 3. Given two classification methods that do not have any hyperparameter to tune, and a training set of *n* examples: (a) describe how to accurately estimate which provides higher predictive accuracy for predictions on new data not in the training set; and (b) what assumptions are made about the training data for this to work.

Solution:

(a) We can use K-fold cross-validation. First randomly divide the dataset into K folds. Then apply the two classification algorithms to train on the K-1 fold and test on remaining fold (validation set). Then average the accuracy over the K folds, and the algorithm with the higher average accuracy is better.

(b). The examples in the training dataset are drawn iid.

• Parametric models

Question 0. TRUE or FALSE

A classification method with low bias and more variance should always be preferred over one with more bias and lower variance.

FALSE. We prefer low bias and low variance. So it is not clear which one in the question is more preferred. The mean squared error is bias^2 + variance, which we should compare on.

Question [4.1] (see ex4_1.m in tut_1_code.zip on Piazza under Resources -> Homework)

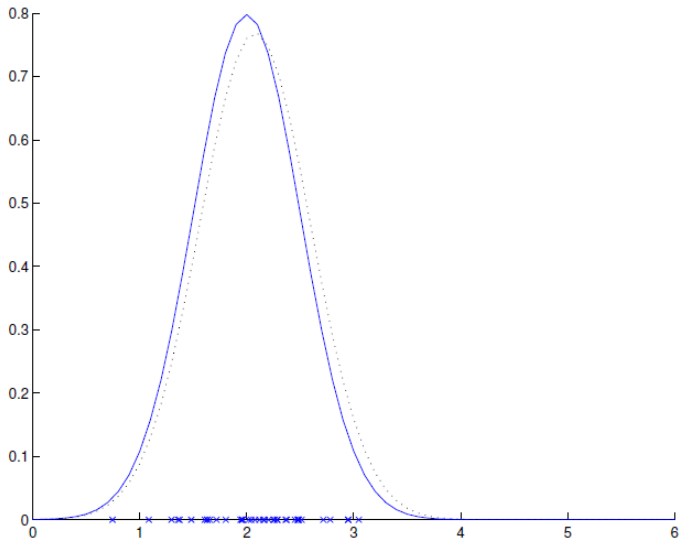
Question [4.2] (you need to understand it, but deriving it by yourself is optional)

We add the constraint as a Lagrange term and maximize it:

$$
\begin{aligned}
J(p_i) &= \sum_i \sum_t x_i^t \log p_i + \lambda(1 - \sum_i p_i) \\
\frac{\partial J}{\partial p_i} &= \frac{\sum_t x_i^t}{p_i} - \lambda = 0 \\
\lambda &= \frac{\sum_t x_i^t}{p_i} \Rightarrow p_i \lambda = \sum_t x_i^t \\
\sum_i p_i \lambda &= \sum_i \sum_t x_i^t \Rightarrow \lambda = \sum_t \sum_i x_i^t \\
p_i &= \frac{\sum_t x_i^t}{\sum_t \sum_i x_i^t} = \frac{\sum_t x_i^t}{N} \text{ because } \sum_i x_i^t = 1
\end{aligned}
$$

Question [4.3]

The Matlab code ex4_3.m in tut_1_code.zip on Blackboard under Tutorials, and below is an example output.



Actual Gaussian density (continuous line),

a sample drawn from it (×),

and a Gaussian fitted to the sample (dotted line).

Question [4.4-4.5, 4.10]. Solution is available in the textbook.

Question [4.9].

Taking any instance has less bias than taking a constant but has higher variance. It has higher variance than the average and it may have higher bias. If the sample is ordered so that the instance we pick is the minimum, variance decreases (different minima tend to get more similar to each other) and bias may also increase. The Matlab code is given in `ex4_9.m` and running it, we get the plot of figure 4.5 with the following output:

```
Order 0 B=1.87 V=0.10 E=1.97
First inst B=2.63 V=2.30 E=4.93
Min inst B=14.42 V=0.53 E=14.95
```

Question [4.6].

The Matlab code is given in ex4_6.m, and its output plot is below.

Output:

```
Real values C1:(3.0,1.0)- C2:(2.0,0.3)
Estimates C1:(3.45,1.23)- C2:(2.15,0.26)
Actual intersect pts: 2.49 1.31
Estimated intersect pts: 2.64 1.53
```

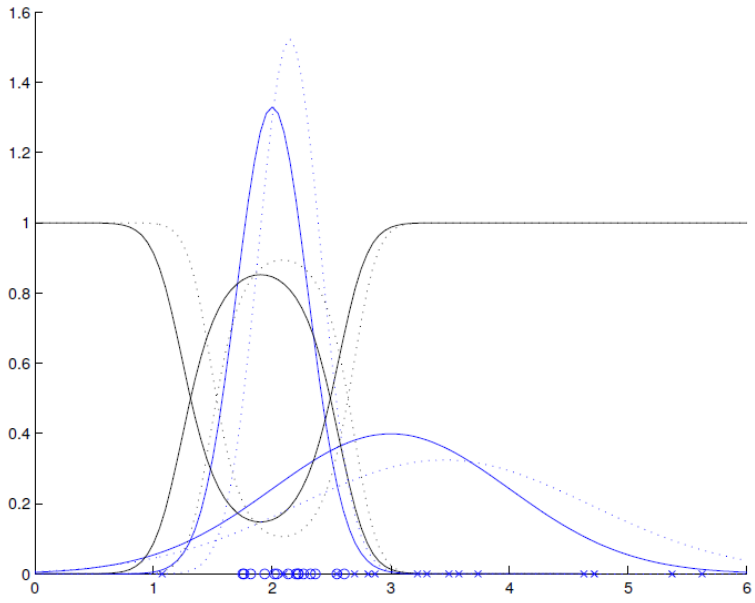


Figure 4.2 Actual densities, posteriors, drawn samples and the fitted densities and posteriors for a two-class problem. The continuous lines are the actual densities and posteriors and the dotted lines are the estimations using the data points shown by '×' and '∘'.

4.7 The Matlab code given in `ex4_7.m` samples data from $2\sin(1.5x) + \mathcal{N}(0,1)$, and fits five times polynomials of order 1, 2 and 3. It also calculates their bias and variance. This is the code used to generate figure 4.5 of the book (with orders 1, 3 and 5). The plots for orders 1, 2, and 3 are given below

Output:

```
Order 1 B=1.80 V=0.29 E=2.09
Order 2 B=1.25 V=0.38 E=1.63
Order 3 B=0.28 V=0.30 E=0.58
```

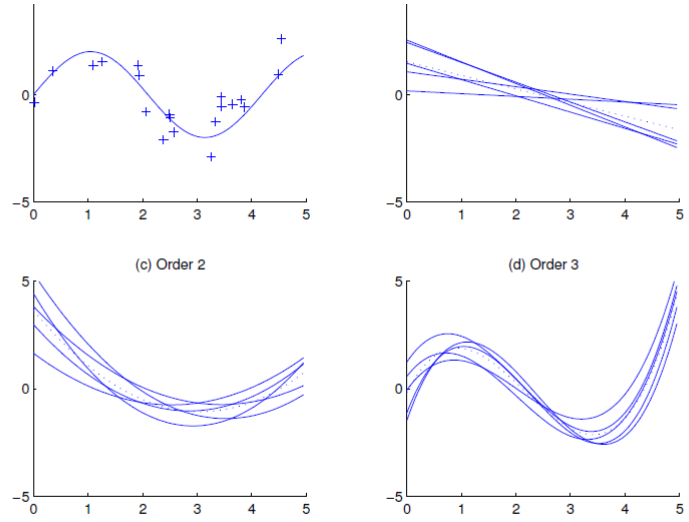


Figure 4.3 Function, one random noisy sample and five fitted polynomials of order 1, 2 and 3.

4.8 With small datasets and complex models, the fit will have high variance because there is not enough data to sufficiently constrain the model and the fit will follow noise. In such a case, it is better to use a simple model though it risks having high bias.
If we run the program of exercise 4.7 here with six data points (`ex4_8.m` has $N = 6$ on line 2), we get the result in the figure below and the following output:

```
Order 1 B=1.66 V=0.37 E=2.03
Order 2 B=1.31 V=1.59 E=2.90
Order 3 B=3.74 V=2.86 E=6.60
```

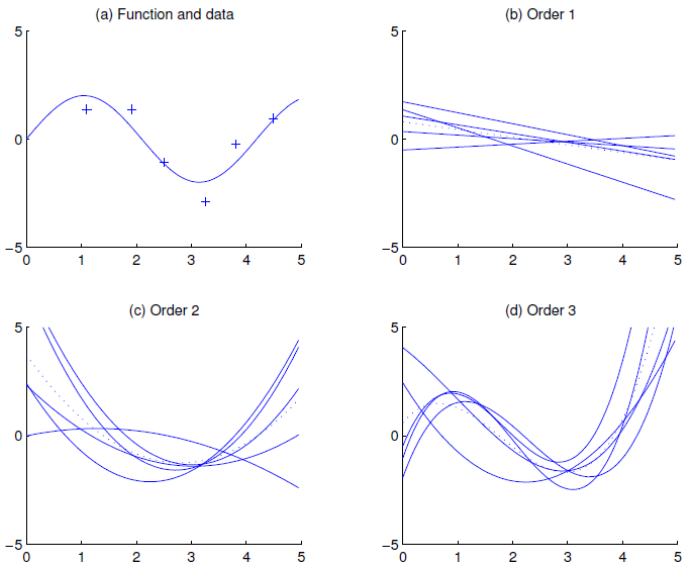


Figure 4.4 Function, one random noisy sample and five fitted polynomials of order 1, 2 and 3.

Question 0. TRUE or FALSE.

a) The MAP estimate and the maximum likelihood estimate converge to the same solution given infinite data and a reasonable prior distribution (providing non-zero probability everywhere).

TRUE

b) Maximum a posteriori (MAP) estimation averages over the posterior parameter distribution to make predictions for new data.

FALSE (MAP is point estimation. Only Bayesian estimation does the above)

Question 1. Multiple choice

a) Consider maximum likelihood estimation (MLE), maximum a posteriori estimation (MAP), and Bayesian estimation (Bayes) with a prior with non-zero probability for all model parameters. Which of the following are true?
(a) MLE and MAP with a uniform prior are equivalent.
(b) MLE and Bayes with a uniform prior are equivalent.
(c) With infinite amounts of data, MLE and Bayes will converge to the same estimates.
(d) With infinite amounts of data, MLE and MAP will converge to the same estimates.

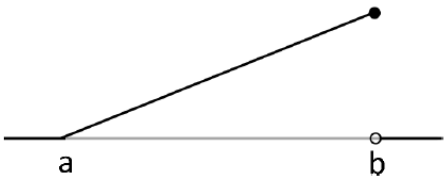
Solution: (a)(d).

(b) is incorrect because with a uniform prior, both MLE and MAP are finding the mode of the posterior distribution $P(\theta|X)$, while Bayesian estimation is finding its mean. Of course the mean and mode of a distribution can be different, e.g., exponential distribution.

(c) is incorrect because with infinite amount of data, both MLE and MAP are finding the mode of the posterior distribution $P(\theta|X)$, while Bayesian estimation is finding its mean. Again, like in (b), they can be very different.

Question 2.

Consider the "ramp" continuous probability distribution. It is defined by two parameters, a and b. For $x < a$, the probability density is 0. Between a and b, the probability density increases with a fixed slope until it is maximized at b. For $x > b$, the probability density is again 0.



a) What is the probability density function of this distribution at $x = b$ (as a function of a and b)?

$f(x = b) =$

b) Given two datapoints, x_1 and x_2, what is the maximum likelihood estimate for b? (Hint: no calculus is required.)

$\hat{b} =$

c) If we estimated using mean of the Bayesian posterior and a reason-able prior, would this Bayesian mean estimate of b be SMALLER than the MLE estimate, THE SAME as the MLE estimate, or LARGER than the MLE estimate? Why?

d) Given this maximum likelihood estimate for b, what is the likelihood function given the pair of datapoints, $P(x_1, x_2|a, b)$?

e) What is the maximum likelihood estimate for parameter a given the same two datapoints x_1 and x_2? (Hint: calculus is required and logarithms may be useful.)

Solution:

(a) $f(x = b) = \frac{2}{b-a}$ (to make the integral of pdf be 1)

(b) $\hat{b} = \max(x_1, x_2)$

To maximize the likelihood, we need to ensure that x_1, x_2 are both in $(a, b]$ so that $p(x_1)$ and $p(x_2)$ are both positive. That is, $b >= \max(x_1, x_2)$. For all such b, the likelihood decreases as b grows. So the MLE of b is $\max(x_1, x_2)$.

(c) Bayesian mean estimate of b will be LARGER than MLE.
MLE sets b to $\max(x_1, x_2)$.

As for Bayesian estimation, the posterior of b has 0 density for $b < \max(x_1, x_2)$ because the likelihood is 0 for such b.

So the mean of the posterior of b must be greater than $\max(x_1, x_2)$.

(d) By property of triangle, we get

$$P(x) = \begin{cases} \frac{2(x-a)}{(\hat{b}-a)^2} & \text{if } x \in [a, \hat{b}] \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned} P(x_1, x_2|a, b) &= P(x_1|a, b)P(x_2|a, b) \\ &= \frac{2(\min(x_1, x_2) - a)}{(\max(x_1, x_2) - a)^2} * \frac{2}{\max(x_1, x_2) - a} \end{aligned}$$

(e) Let $\nabla_a \log P(x_1, x_2|a, b) = 0$, we get $\hat{a} = \frac{3\min(x_1, x_2) - \max(x_1, x_2)}{2}$

Question 3. Consider binary-valued variable $x \in \{0, 1\}$, distributed according to a Bernoulli distribution

$$f_\lambda(x) = \lambda^x(1 - \lambda)^{1-x}.$$

Note $f_\lambda(x)$ is effectively $P(x|\lambda)$, the likelihood probability.

Suppose the prior distribution is a distribution over $\lambda \in [0, 1]$ with

$$P\left(\lambda = \frac{1}{4}\right) = \frac{1}{4}, \quad P\left(\lambda = \frac{1}{2}\right) = \frac{1}{2}, \quad P\left(\lambda = \frac{3}{4}\right) = \frac{1}{4}.$$

That is, $P(\lambda) = 0$ for all other values of λ.

Now suppose we have observed three data points $x_1 = 1$, $x_2 = 1$, and $x_3 = 0$.

(a) What is the maximum likelihood estimate (MLE) for λ, given the dataset $D = \{x_1, x_2, x_3\}$?

(b) What is the Bayesian posterior probability $P(\lambda|D)$ for each value of $\lambda = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}$?

(c) What is the maximum a posteriori (MAP) estimate for λ given the dataset *D* and prior probability distribution?

(d) What is the posterior probability of a new data point x_4 under the fully Bayesian estimation of λ?

(e) What is the posterior probability of a new data point x_4 under the MAP estimation of λ?

Solutions:

(a) The likelihood of the dataset *D* given θ is:

$$L = P(D|\lambda) = P(x_1|\lambda)P(x_2|\lambda)P(x_3|\lambda) = f_\lambda(x_1)f_\lambda(x_2)f_\lambda(x_3)$$
$$= \lambda^1(1-\lambda)^{1-1} \cdot \lambda^1(1-\lambda)^{1-1} \cdot \lambda^0(1-\lambda)^{1-0}$$
$$= \lambda^2(1-\lambda).$$

To maximize *L* over $\lambda \in [0, 1]$, we get the MLE as $\lambda = 2/3$. Note that $P(\lambda = 2/3)$ is 0 according to the prior probability. But this does not matter because MLE does not care about the prior distribution.

(b) The posterior distribution is

$$P(\lambda|D) = \frac{P(D|\lambda)P(\lambda)}{P(D)} \propto P(D|\lambda)P(\lambda) = \lambda^2(1-\lambda)P(\lambda),$$

where the last equality is from (a). Now plug in the value of $P(\lambda)$

$$\lambda^2(1-\lambda)P(\lambda) \quad = \quad \begin{cases} \frac{1}{4^2}(1-\frac{1}{4})\frac{1}{4} & if\ \lambda = \frac{1}{4} \\ \frac{1}{2^2}(1-\frac{1}{2})\frac{1}{2} & if\ \lambda = \frac{1}{2} \\ \left(\frac{3}{4}\right)^2(1-\frac{3}{4})\frac{1}{4} & if\ \lambda = \frac{3}{4} \end{cases} \quad = \quad \begin{cases} 0.0117 & if\ \lambda = \frac{1}{4} \\ 0.0625 & if\ \lambda = \frac{1}{2} \\ 0.0352 & if\ \lambda = \frac{3}{4} \end{cases} \quad (*)$$

Summing up $0.0117 + 0.0625 + 0.0352 = 0.1094$. Therefore, we can normalize by

$$P(\lambda|D) \quad = \quad \begin{cases} \frac{1}{0.1094} \cdot 0.0117 & if\ \lambda = \frac{1}{4} \\ \frac{1}{0.1094} \cdot 0.0625 & if\ \lambda = \frac{1}{2} \\ \frac{1}{0.1094} \cdot 0.0352 & if\ \lambda = \frac{3}{4} \end{cases} \quad = \quad \begin{cases} 0.1071 & if\ \lambda = \frac{1}{4} \\ 0.5714 & if\ \lambda = \frac{1}{2} \\ 0.3214 & if\ \lambda = \frac{3}{4} \end{cases} \quad (**)$$

(c) Based on (*), we can see that $\lambda = \frac{1}{2}$ attains the highest value of $P(\lambda|D)$. Therefore, the MAP estimate is $\lambda^{MAP} = \frac{1}{2}$. Note we can see this before performing normalization, because all the three last quantities in (*) will be divided by the same normalizer (0.1094), and that does not change which λ produces the highest value.

(d) Based on (b), we obtain

$$
\begin{aligned}
&P(x_4|D) \\
&= P\left(x_4\middle|\lambda=\frac{1}{4}\right)P\left(\lambda=\frac{1}{4}\middle|D\right)+P\left(x_4\middle|\lambda=\frac{1}{2}\right)P\left(\lambda=\frac{1}{2}\middle|D\right)+P\left(x_4\middle|\lambda=\frac{3}{4}\right)P\left(\lambda=\frac{3}{4}\middle|D\right) \\
&= \left(\frac{1}{4}\right)^{x_4}\left(1-\frac{1}{4}\right)^{1-x_4}0.1071+\left(\frac{1}{2}\right)^{x_4}\left(1-\frac{1}{2}\right)^{1-x_4}0.5714+\left(\frac{3}{4}\right)^{x_4}\left(1-\frac{3}{4}\right)^{1-x_4}0.3214 \\
&= \begin{cases} 0.4464 & if\ x_4=0 \\ 0.5536 & if\ x_4=1 \end{cases}.
\end{aligned}
$$

Note here we need to use the normalized value of $P(\lambda|D)$ in (**).

(e) Based on (c) which shows $\lambda^{MAP}=\frac{1}{2}$, we have

$$
P(x_4|\lambda^{MAP)}=P\left(x_4\middle|\frac{1}{2}\right)=f_{\frac{1}{2}}(x_4)=\left(\frac{1}{2}\right)^{x_4}\left(1-\frac{1}{2}\right)^{1-x_4}=\begin{cases}0.5 & if\ x_4=0\\ 0.5 & if\ x_4=1\end{cases}.
$$