# 6.867: Exercises (Week 2)

Sept 8, 2016

## Contents

# 1 One parameter, two estimators

In this problem, we're going to explore the bias-variance trade-off in a very simple setting. We have a set of unidimensional data, $x^{(1)}, \ldots, x^{(n)}$, drawn from the positive reals. Consider a simple model for its distribution (in a later problem we will consider a slightly different model):

- **Model 1:** The data are drawn from a uniform distribution on the interval $[0, b]$. This model has a single positive real parameter $b$, such that $0 < b$.

We are interested in estimates of the mean of the distribution.

(a) What's the mean of the Model 1 distribution?

Let's start by considering the situation in which the data were, in fact, drawn from an instance of the model under consideration: a uniform distribution on $[0, b]$ (for model 1),

In model 1, the ML estimator for $b$ is $b_{\mathbf{ml}} = \max_i x^{(i)}$. The likelihood of the data is:

$$L(b_{\mathbf{ml}}) = \prod_{i=1}^{n} \begin{cases} b_{\mathbf{ml}}^{-1} & \text{if } x^{(i)} \leqslant b_{\mathbf{ml}} \\ 0 & \text{otherwise} \end{cases}$$

We can see that if $b_{\mathbf{ml}} < x^{(i)}$, for any $x^{(i)}$, then the likelihood of the whole data set must be 0. So, we should pick $b_{\mathbf{ml}}$ to be as small as possible subject to the constraint that $b_{\mathbf{ml}} \geqslant x^{(i)}$, which means $b_{\mathbf{ml}} = \max_i x^{(i)}$.

To understand the properties of this estimator we have to start by deriving their PDFs. The minimum and maximum of a data set are also known as their first and nth *order statistics*, and sometimes written $x^{[1]}$ and $x^{[n]}$ (we're using square brackets to distinguish these from our notation for samples in a data set).

In model 1, we just need to consider the distribution of $b_{\mathbf{ml}}$. Generally speaking, the pdf of the maximum of a set of data drawn from pdf $f$, with cdf $F$, is:

$$f_{b_{\mathbf{ml}}}(x) = nF(x)^{n-1} f(x) \tag{1}$$

The idea is that, if $x$ is the maximum, then $n-1$ of the other data values will have to be less than $x$, and the probability of that is $F(x)^{n-1}$, and then one value will have to equal $x$, the probability of which is $f(x)$. We multiply by $n$ because there are $n$ different ways to choose the data value that could be the maximum.

(b) What is the maximum likelihood estimate of the mean, $\mu_{\mathbf{ml}}$, of the distribution?

(c) What is $f_{b_{\mathbf{ml}}}$ for this particular case where the data are drawn uniformly from 0 to $b$?

(d) Write an expression for the expected value of $\mu_{\mathbf{ml}}$, as an integral,

In fact, there's a nice closed form expression, which you can use in the following questions:

$$E[\mu_{\mathbf{ml}}] = \frac{b}{2} \frac{n}{(n+1)} \quad .$$

(e) What is the squared bias of $\mu_{\mathbf{ml}}$? Is this estimator unbiased? Is it asymptotically unbiased? (Reminder: $\text{bias}^2(\theta_{\mathbf{ml}}) = (\mathsf{E}_D[\theta_{\mathbf{ml}}] - \theta)^2$.)

(f) Write an expression for the variance of $\mu_{\mathbf{ml}}$, as an integral.

The closed form for the variance is

$$V[\mu_{\mathbf{ml}}] = \frac{b^2}{4}\frac{n}{(n+1)^2(n+2)} \quad .$$

(g) What is the mean squared error of $\mu_{\mathbf{ml}}$? (Reminder: $\text{MSE}(\theta_{\mathbf{ml}}) = \text{bias}^2(\theta_{\mathbf{ml}}) + \text{var}(\theta_{\mathbf{ml}})$.)

(h) So far, we have been considering the error of the *estimator*, comparing the estimated value of the mean with its actual value. We will often want to use the estimator to make predictions, and so we might be interested in the expected error of a prediction.

Assume the loss function for your predictions is $L(g, a) = (g - a)^2$. Given an estimate $\hat{\mu}$ of the mean of the distribution, what value should you predict?

What is the expected loss (risk) of this prediction? Take into account both the error due to inaccuracies in estimating the mean as well as the error due to noise in the generation of the actual value.

We might consider something other than the MLE for Model 1 (labeled o for other). Consider the estimator

$$\mu_o = \frac{x^{[n]}(n+1)}{2n} \quad .$$

where $x^{[n]}$ is the maximum of the data set.

(i) Write an expression for the expected value of this version of $\mu_o$ as an integral. Then solve the integral.

(j) What is the squared bias of this estimator for $\mu_o$? Is this estimator unbiased? Is it asymptotically unbiased?

(k) Write an expression for the variance of $\mu_o$ as an integral.

The closed form for the variance is

$$V[\mu_o] = \frac{b^2}{4n(n+2)} \quad .$$

(l) What is the mean squared error of this version of $\mu_o$?

(m) What are the relative advantages of the estimator from the previous question and this one?

## 2 One problem, two models

In this problem, we're going to continue exploring the bias-variance trade-off in a very simple setting. We have a set of unidimensional data, $x^{(1)}, \ldots, x^{(n)}$, drawn from the positive reals. We will consider two different models for its distribution:

- **Model 1:** The data are drawn from a uniform distribution on the interval $[0, b]$. This model has a single positive real parameter $b$, such that $0 < b$.
- **Model 2:** The data are drawn from a uniform distribution on the interval $[a, b]$. This model has two positive real parameters, $a$ and $b$, such that $0 < a < b$.

We are interested in comparing estimates of the mean of the distribution, derived from each of these two models.

### 2.1 Using Model 2

(a) What's the mean of the Model 2 distribution?

(b) Let's consider the situation in which the data were, in fact, drawn from an instance of the model under consideration: either a uniform distribution on $[0, b]$ (for model 1) or a uniform distribution on $[a, b]$ (for model 2).

In model 1, the ML estimator for $b$ is $b_{\mathbf{ml}} = \max_i x^{(i)}$. The likelihood of the data is:

$$L(b_{\mathbf{ml}}) = \prod_{i=1}^{n} \begin{cases} b_{\mathbf{ml}}^{-1} & \text{if } x^{(i)} \leqslant b_{\mathbf{ml}} \\ 0 & \text{otherwise} \end{cases}$$

We can see that if $b_{\mathbf{ml}} < x^{(i)}$, for any $x^{(i)}$, then the likelihood of the whole data set must be 0. So, we should pick $b_{\mathbf{ml}}$ to be as small as possible subject to the constraint that $b_{\mathbf{ml}} \geqslant x^{(i)}\ \forall i$, which means $b_{\mathbf{ml}} = \max_i x^{(i)}$.

By a similar argument in model 2, the ML estimator for $b$ remains the same and the ML estimator for $a$ is $a_{\mathbf{ml}} = \min_i x^{(i)}$. To understand the properties of these estimators we have to start by deriving their PDFs. The minimum and maximum of a data set are also known as their first and nth *order statistics*, and sometimes written $x^{(1)}$ and $x^{(n)}$.

We started our analysis of Model 1 in question 1. Now, let's do the same thing, but for the MLE for model 2. We have to start by thinking about the joint distribution of MLE's $a_{\mathbf{ml}}$ and $b_{\mathbf{ml}}$. Generally speaking, the joint pdf of the minimum and the maximum of a set of data drawn from pdf $f$, with cdf $F$, is

$$f_{a_{\mathbf{ml}}, b_{\mathbf{ml}}}(x, y) = n(n-1)(F(y) - F(x))^{n-2} f(x) f(y) \ .$$

Explain in words why this makes sense.

What is $f_{a_{\mathbf{ml}}, b_{\mathbf{ml}}}$ in the particular case where the data are drawn uniformly from $a$ to $b$?

Write an expression for the expected value of $\mu_{\mathbf{ml}}$ in terms of an integral.

Here's what it should integrate to:

$$E[\mu_{\mathbf{ml}}] = \frac{a+b}{2} \ .$$

(c) What is the squared bias of $\mu_{\mathbf{ml}}$? Is this estimator unbiased? Is it asymptotically unbiased?

(d) Write an expression for the variance of $\mu_{\mathbf{ml}}$ in terms of an integral.

The closed form for the variance is

$$V[\mu_{\mathbf{ml}}] = \frac{(b-a)^2}{2(n+1)(n+2)} \quad .$$

(e) What is the mean squared error of $\mu_{\mathbf{ml}}$?

## 2.2 Comparing Models

What if we have data that is actually drawn from the interval $[0, 1]$? Both models seem like reasonable choices.

(a) Show plots that compare the bias, variance, and MSE of each of the estimators we've considered on that data, as a function of $n$. (Use the formulas above; don't do it by actually generating data). Write a paragraph in English explaining your results. What estimator would you use?

(b) Now, what if we have data that is actually drawn from the interval $[.1, 1]$? It seems like model 2 is the only reasonable choice. But is it?

We already know the bias, variance, and MSE for model 2 in this case. But what about the MLE and unbiased estimators for model 1? Let's characterize the general behavior when we use the estimator $\mu_{\mathbf{ml}} = x^{(n)}(n+1)/(2n)$ on data drawn from an interval $[a, b]$.

Write an expression for the expected value of $\mu_{\mathbf{ml}}$ in terms of an integral.

(c) The closed form expression is

$$E[\mu_{\mathbf{ml}}] = \frac{a+bn}{2n} \quad .$$

Explain in English why this answer makes sense.

(d) What is the squared bias of this $\mu_{\mathbf{ml}}$? Explain in English why your answer makes sense. Consider how it behaves as $a$ increases, and how it behaves as $n$ increases.

(e) Write an expression for the variance of this $\mu_{\mathbf{ml}}$ in terms of an integral.

The closed form for the variance is

$$V[\mu_{\mathbf{ml}}] = \frac{(b-a)^2}{4n(n+2)} \quad .$$

To save you some tedious algebra, we'll tell you that the mean squared error of this $\mu_{\mathbf{ml}}$ is (apologies for the ugliness; let us know if you find a beautiful rewrite)

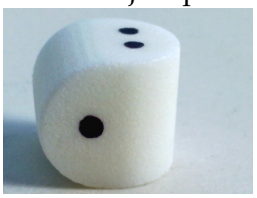
$$\frac{b^2n - 2abn + a^2(2 - 2n + n^3)}{4n^2(n+2)} \quad .$$

(f) Show plots that compare the bias, variance, and MSE of this estimator with the regular model 2 estimator on data drawn from $[0.1, 1]$, as a function of $n$. Are there circumstances in which it would be better to use this estimator? If so, what are they and why? If not, why not?

(g) Show plots of MSE of both estimators, as a function of $n$ on data drawn from $[.01, 1]$ and on data drawn from $[.2, 1]$. How do things change? Explain why this makes sense.

## 3 Which dice factory?

You have just purchased a two-sided die, which can come up either 1 or 2:



You want to use your crazy die in some betting games with friends later this evening, but first you want to know the probability that it will roll a 1.

You know know it came either from factory 0 or factory 1, but not which.

Factory 0 produces dice that roll a 1 with probability $\phi_0$. Factory 1 produces dice that roll a 1 with probability $\phi_1$. You believe initially that with probability $\eta_0$ that it came from factory 1.

(a) Without seeing any rolls of this die, what would be your predicted probability that it would roll at 1?

(b) If we roll the die and observe the outcome, what can we infer about where the coin was manufactured?

(c) More concretely, let's assume that:

- $\phi_0 = 1$: dice from factor 0 always roll a 1
- $\phi_1 = 0.5$: dice from factory 1 are fair (roll at 1 with probability 0.5)
- $\eta_0 = 0.7$: we think with probability 0.7 that this die came from factory 1

Now we roll it, and it comes up 1! What is your posterior distribution on which factory it came from? What is your predictive distribution on the value of the next roll?

(d) You roll it again, and it comes up 1 again.

Now, what is your posterior distribution on which factory it came from? What is your predictive distribution on the value of the next roll?

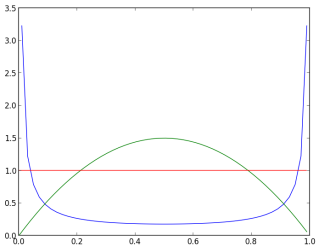
(e) Instead, what if it rolls a 2 on the second roll?

(f) In the general case (not using the numerical values we have been using) prove that if you have two observations, and you use them to update your prior in two steps (first conditioning on one observation and then conditioning on the second), that no matter which order you do the updates in you will get the same result.

## 4 Beta-Binomial

(a) Label which of the lines in the figure below correspond to:

1. Beta(0.1, 0.1)
2. Beta(1,1)
3. Beta(2,2)




We are estimating the probability that a coin comes up heads.

(b) What does it mean to have a prior of Beta$(2, 2)$?

(c) If that's the prior, what is the posterior after seeing 3 heads and 2 tails?

(d) What are the mean and mode of that posterior?

(e) What does it mean to have a prior of Beta$(2, 3)$?

(f) If that's the prior, what is the posterior after seeing 3 heads and 2 tails?

(g) What are the mean and mode of that posterior?

## 5 Emergency Room, Reconsidered

You are a young doctor, working off your federal medical school tuition grant in southern North Dakota. It's your fourth day on the job. You are all alone in the emergency room (ER) when Pat comes in complaining of chest pain.

You have to predict whether Pat is having a heart attack (H) or indigestion (I). Your loss function is:

$$L(g, a) = \begin{cases} 0 & \text{if } g = a \\ 1 & \text{if } g = \text{"H" and } a = \text{"I"} \\ 10 & \text{if } g = \text{"I" and } a = \text{"H"} \end{cases}$$

You have seen three previous patients who exhibited chest pain, none of whom were actually having a heart attack.

(a) You use those three data points to make a point estimate of the probability that Pat is having a heart attack and then use it to make the prediction that minimizes the empirical risk. What do you predict? What is the empirical risk of that prediction?

> Do you think the empirical risk of this predictor is a good measure of how useful it will be?

(b) The next morning, you think more carefully and decide it would be better to forget all your previous experience and simply view each new patient with an open mind. So, you use some ideas from this week's lectures. Let Q be a random variable representing the probability that

a random patient walking into your ER will be having a heart attack. You have a uniform prior on Q.

What is the prediction that minimizes risk for a random patient walking into your ER? What is the risk of that prediction?

(c) Later that afternoon, you figure it would be better to combine approaches. So, what if you started with a uniform prior, but then observed three patients all of whom had indigestion?

What would be your posterior distribution on Q? What prediction should you make? What is the risk (under the posterior distribution) of that prediction?

(d) That evening, really worried that you haven't had enough experience in these matters, and beginning to question your judgment about accepting this job, you decide to call your friend Chris who is working at Mass General. Chris has seen 20 patients with indigestion and 1 with heart attack. You use Chris's experience to construct a prior distribution, and then update it with your own (3 patients with indigestion).

What would be your posterior distribution on Q? What prediction should you make? What is the risk (under the posterior distribution) of that prediction?

(e) At 2AM, questioning the meaning of life, you are quite sure that you should have become a poet. You are so uncertain of your ability to make predictions that you call your former professor who is the head of the emergency medicine department at Gotham City Hospital. Herr Prof. Dr. Strangelove has seen 2000 patients with indigestion and 20 with heart attack. You use Dr. Strangelove's experience to construct a prior distribution, and then update it with your own (3 patients with indigestion).

What would be your posterior distribution on Q? What prediction should you make? What is the risk of that prediction?

Is there a potential problem with using Dr. Strangelove's data to help construct your prior?

## 6 Abby Normal

Dr. Frahnkensteen is designing an artificial cranium, but he needs to know how big to make it; his design goal is to be a good fit to 80% of brains. So, he wants to get a good estimate of the distribution of the sizes of brains in the local population. Since brains are kind of squishy, we will just consider the total volume of the brain, a one-dimensional quantity.

The Dr. has considerable previous experience with brains and thinks their distribution is well modeled as a Gaussian distribution with with a variance of 75cc. But he's not at all sure about the mean of this current population. He thinks it might be somewhere around 1100cc.

(a) One way to express the Dr.'s uncertainty about the distribution of brain sizes in his local population is to put a Gaussian distribution *on the mean* of the local distribution.

What are the hyper-parameters of this distribution? Pick some to model Dr. F's situation (they're not completely determined by the story).

(b) Dr. F. sends his assistant Eygor out to get a new brain from the local population. Eygor brings back one that is 1500cc! What should the posterior be?

Start by solving this problem algebraically. Write down the prior and the observation likelihood function symbolically. Then, derive a form for the posterior.

What actual numerical values do you get, given your answer to the previous question, and the observation of 1500cc?

(c) How is the new mean related to the old mean and the observation?

(d) What can we say about how the variance behaves when an observation is made?

(e) What is Dr. F's. posterior predictive distribution? First find it symbolically, then numerically.

(f) If Eygor brought back 10 more brains from the local morgue, would Dr. F. be able to update his prior in some way that is more efficient than doing the individual update procedure 10 times?

## 7 Copy that

You have just bought a copy machine at a garage sale. You know it is one of two possible models, $m_1$ or $m_2$, but the tag has fallen off, so you're not sure which.

You do know that $m_1$ machines have a 0.1 "error" (bad copy) rate and $m_2$ machines have a 0.2 error rate.

(a) You use your machine to make 1000 copies, and 140 of them are bad. What is the maximum likelihood estimate of the machine's error rate? Explain why. (Remember that you're sure it's one of those two types of machines).

(b) Looking more closely, you can see part of the label, and so you think that, just based on the label it has a probability 0.2 of being an $m_1$ type machine and a probability 0.8 of being an $m_2$ type machine. If you take that to be your prior, and incorporate the data from part a, what is your posterior distribution on the type of the machine?

(c) Given that posterior, what is the probability that the next copy will be a failure?

(d) You intend to sell this machine on the web. Because it's used, you have to sell it with a warranty. You can offer a gold or a silver warranty. If it has a gold warranty and the buyer runs it for 1000 copies and gets more than 150 bad copies, then you are obliged to pay \$1000 in damages; if it has a silver warranty, you have to pay damages if it generates more than 300 bad copies in 1000 copies. Your maximum reasonable asking price for a machine with a gold warranty is \$300; for a machine with a silver warranty, it is \$100. You can assume the machine will sell at these prices. What type of warranty should you offer on this machine?

(e) Under what conditions would it be better to just throw the machine away, rather than try to sell it?

## 8 Bayes (Bishop 2.7)

Consider a binomial random variable $x$ given by:

$$\binom{N}{m}\mu^m(1-\mu)^{N-m} \quad (2.9)$$

with prior distribution for $\mu$ given by the beta distribution:

$$\text{Beta}(\mu; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \quad (2.13)$$

and suppose we have observed $m$ occurrences of $x = 1$ and $l$ occurrences of $x = 0$. Show that the posterior mean value of $\mu$ lies between the prior mean and the maximum likelihood estimate for $\mu$.

To do this, show that the posterior mean can be written as $\lambda$ times the prior mean plus $(1-\lambda)$ times the maximum likelihood estimate, where $0 \leqslant \lambda \leqslant 1$. This illustrates the concept of the posterior distribution being a compromise between the prior distribution and the maximum likelihood solution.

## 9 Dirichlet Priors

*Exercise borrowed from Stat180 at UCLA. See Bishop, sections 2.1 and 2.2 for background on Beta and Dirichlet distributions.*

The Dirichlet distribution is a multivariate version of the Beta distribution. When we have a coin with two outcomes, we really only need a single parameter $\theta$ to model the probability of heads. But now let's consider a "thick" coin that has three possible outcomes: heads, tails, and edge. Now we need two parameters: $\theta_h$ is the probability of heads, $\theta_t$ is the probability of tails, and then the probability of an edge is $1 - \theta_h - \theta_t$.

The random variables $(V, W) \in [0, 1]$ and such that $V + W \leqslant 1$ have a Dirichlet distribution with parameters $\alpha_1, \alpha_2, \alpha_3$ if their joint density is

$$f(v, w) = v^{\alpha_1-1}w^{\alpha_2-1}(1-v-w)^{\alpha_3-1}\frac{\Gamma(\alpha_1+\alpha_2+\alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} \quad .$$

This is a direct generalization of the Beta distribution. (Note that $\Gamma$ refers to the Gamma function, which is a generalization of factorial.)

(a) If $(\theta_h, \theta_t)$ have a Dirichlet distribution as above, what is the marginal distribution of $\theta_h$?

(b) Suppose you are playing with a thick coin, and get results $x^{(1)} \ldots x^{(n)}$, resulting in $H$ heads and $T$ tails out of $n$ throws. Given $\theta_h$ and $\theta_t$ the random variables $H$ and $T$ have a multinomial distribution:

$$\Pr(H, T|\theta_h, \theta_t) = \frac{n!}{H!T!(n-H-T)!}\theta_h^H\theta_t^T(1-\theta_h-\theta_t)^{n-H-T} \quad .$$

Assume a uniform prior on the space of possible values of $\theta_h$ and $\theta_t$ (remembering that they are constrained such that $\theta_h \geqslant 0$, $\theta_t \geqslant 0$, and $\theta_h + \theta_t \leqslant 1$). What is the posterior distribution for $\theta_h$ and $\theta_t$?

(c) In this same setting, what is the predictive distribution for getting another head? That is, what's $\Pr(x^{(n+1)} = \text{heads} \mid x^{(1)} \dots x^{(n)})$?

(d) Now assume a Dirichlet prior for $\theta_h$ and $\theta_t$ with parameters $\alpha_1, \alpha_2, \alpha_3$. What is the posterior in this case?

(e) In this same case, what is the predictive distribution?

(f) If you assume a squared-error loss on the predicted parameter, that is,

$$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2 \ ,$$

what is the Bayes-optimal estimate of $\theta_h$ and $\theta_t$?

(g) As $n \to \infty$, how do optimal estimates relate to the maximum likelihood estimates and to the prior?

## 10 Parameter estimation

Given a parameterized family of probability models $\Pr(x \mid \theta)$ and a data set $D = (x^{(1)}, \dots, x^{(n)})$ comprised of independent samples $x^{(i)} \approx \Pr(x \mid \theta)$, we fit the model to the data so as to maximize the likelihood (or log-likelihood) of all samples. This gives the maximum-likelihood (ML) estimate of the parameters:

$$\hat{\theta}_{ML} = \arg\max_{\theta} \log \Pr(D \mid \theta)$$

This approach does not express any prior bias as to which values of $\theta$ we should prefer when data is limited.

In the sequel, we consider a regularized approach to parameter estimation. Here, we specify a prior model $\Pr(\theta)$ over the set of allowed parameter settings $\Theta$. Given a prior model, we may then employ Bayes' rule to compute the posterior probability of $\theta$ given the observations:

$$\Pr(\theta \mid D) = \frac{\Pr(D \mid \theta) \Pr(\theta)}{\Pr(D)}$$

where

$$\Pr(D) = \int_{\Theta} \Pr(D \mid \theta) \Pr(\theta) d\theta$$

Then, we fit the model to the data by maximizing the (log-) probability of $\theta$ conditioned on the data,

$$\begin{aligned} \hat{\theta}_{MAP} &= \arg\max_{\theta} \log \Pr(\theta \mid D) \\ &= \arg\max_{\theta} \{\log \Pr(D \mid \theta) + \log \Pr(\theta) - \log \Pr(D)\} \\ &= \arg\max_{\theta} \{\log \Pr(D \mid \theta) + \log \Pr(\theta)\} \end{aligned}$$

Note that we have dropped the $-\log \Pr(D)$ term as this does not depend upon $\theta$ and does not affect the parameter estimate. Hence, we do not need to explicitly evaluate the integral in the denominator. This may be viewed as a penalized log-likelihood criterion, i.e. we maximize

$J(\theta) = \log \Pr(D;\theta) + f(\theta)$ subject to the regularization penalty $f(\theta) = \log \Pr(\theta)$. The parameter estimate $\hat{\theta}_{MAP}$ is known as the maximum a posteriori (MAP) estimate.

In this problem you will construct MAP estimates for the probabilities of a (potentially biased) $M$ -sided die, i.e. $x^{(i)} \in \{1,\ldots,M\}$. We consider the fully-parameterized representation $\Pr(x = k) = \theta_k$, where $0 \leqslant \theta_k \leqslant 1$ for $k = 1,\ldots,M$ and $\sum_k \theta_k = 1$. This simple model has many relevant applications.

Consider a document classification task, where we need class-conditional distributions over words in the documents. Suppose we only consider words $1,\ldots,M$ (for relatively large $M$). Each word in the document is assumed to have been drawn at random from the distribution $\Pr(x = k \mid y;\theta) = \theta_{k|y}$, where $\sum_k \theta_{k|y} = 1$ for each class $y$. Thus the selection of words according to the distribution $\theta_{k|y}$ can be interpreted as a (biased) $M$-sided die.

Now, the probability of generating all words $x^{(1)},\ldots,x^{(n)}$ in a document of length $n$ would be

$$\Pr(D \mid y;\theta) = \prod_{i=1}^{n} \Pr(x^{(i)} \mid y;\theta) = \prod_{i=1}^{n} \theta_{x^{(i)}|y}$$

assuming the document belongs to class $y$. Note that this model cares about how many times each word occurs in the document. It is a valid probability model over the set of words in the document.

Since we typically have very few documents per class, it is important to regularize the parameters, i.e., provide a meaningful prior answer to the class conditional distributions.

Let's start by briefly revisiting ML estimation of the (biased) $M$-sided die. Similarly to calculations you have already performed, the ML estimate of the parameter $\theta$ from $n$ samples is given by the empirical distribution:

$$\hat{\theta}_x = \frac{n(x)}{n}$$

where $n(x)$ is the number of times value $x$ occurred in $n$ samples. The count $n(x)$ is also a *sufficient statistic* for $\theta_x$ as it is all we need to know from the available $n$ samples in order to estimate $\theta_x$.

Next, we consider MAP estimation. To do so, we must introduce a prior distribution over the $\theta$'s. A natural choice for this problem is the Dirichlet distribution

$$\Pr(\theta;\beta) = \frac{1}{Z(\beta)} \prod_{k=1}^{M} \theta_k^{\beta_k}$$

with non-negative hyperparameters $\beta = (\beta_k > 0, k = 1,\ldots,M)$ and where $Z(\beta)$ is just the normalization constant (which you saw earlier and which you do not need to evaluate in this problem).

(a) First, consider this prior model (ignoring the data for the moment). What value of $\theta$ is most likely under this prior model? That is, compute

$$\hat{\theta}(\beta) = \arg\max_{\theta} \log \Pr(\theta;\beta)$$

This is the *a priori* estimate of $\theta$ before observing any data.

(b) Next, given the data $D$, compute the MAP estimate of $\theta$ as a function of the hyperparameters $\beta$ and the data $D$ (use the sufficient statistics $n(x)$):

$$\hat{\theta}_{MAP}(D;\beta) = \arg\max_{\theta} \log \Pr(\theta \mid D;\beta)$$

Note that you do not need to calculate $Z(\beta)$ in order to perform this optimization; you can optimize the penalized log-likelihood $J(\theta) = \log \Pr(D \mid \theta) + f(\theta; \beta)$ with a simple penalty function $f(\theta; \beta)$, as discussed above. Thus we do not have to evaluate the full posterior distribution $\Pr(\theta \mid D; \beta)$ in order to perform the regularization.

(c) Show that your MAP estimate may be expressed as a convex combination of the a priori estimate $\hat{\theta}(\beta)$ and the ML estimate $\hat{\theta}_{ML}(D)$. The means that we may write

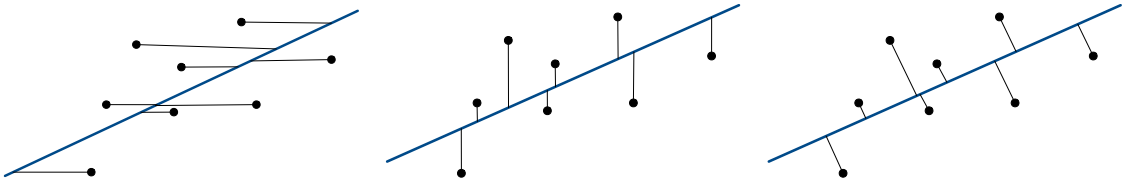
$$\hat{\theta}_{MAP}(D; \beta) = (1 - \lambda)\hat{\theta}_{ML}(D) + \lambda\hat{\theta}(\beta)$$

for some $\lambda \in [0, 1]$. Note that the same convex combination holds for each component $\theta_x$. Determine $\lambda$ as a function of the number of samples $n$ and the hyperparameters $\beta$.

As this shows, one way of thinking of a prior distribution is that it is a proxy for any data we have observed in the past but no longer have available. The normalized parameters $\hat{\beta}_i = \beta_i / N$, where $N = \sum_i \beta_i$, express our prior estimate of the parameters $\theta$ while the normalization parameter $N$ expresses how strongly we believe in that prior estimate.

## 11 Residue removal

Which of the following images shows the error that is minimized in ordinary least-squares regression? Write down the formula and explain how it's related to the small lines in the picture.



## 12 Weighted least squares regression

You are trying to build a predictor with data that you gathered on two different days with two different instruments. We know that data set 1, consisting of $n$ pairs, $(x^{(i)}, y^{(i)})$ has a conditional Gaussian distribution

$$y \sim \text{Normal}(\mathbf{x} \cdot \theta, \sigma_1^2) \ ,$$

and data set 2, consisting of $m$ pairs $(u^{(i)}, v^{(i)})$ has a conditional Gaussian distribution that differs only in the variance:

$$v \sim \text{Normal}(\mathbf{u} \cdot \theta, \sigma_2^2) \ ,$$

The parameter vector $\theta$ and all of the $x^{(i)}$ and $u^{(i)}$ are vectors in $\mathbb{R}^d$, and the $y^{(i)}$ and $v^{(i)}$ are in $\mathbb{R}$.

(a) Derive the maximum-likelihood estimator for $\theta \in \mathbb{R}^d$. You can assume that there is no special $\theta_0$. **We strongly recommend that you do this in matrix-vector form.**

(b) Argue that it makes sense for extreme relative values of $\sigma_1$ and $\sigma_2$.

# 13 Ridge Regression

(a) (Bishop 3.4)
Consider a linear model of the form

$$y(x, \mathbf{w}) = w_o + \sum_{i=1}^{D} w_i x_i \qquad (3.105)$$

together with a sum-of-squares error function of the form

$$Err_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x^{(n)}, \mathbf{w}) - t_n\}^2 \qquad (3.106)$$

where $t_n$ is the true value for $x^{(n)}$. Now suppose that Gaussian noise $\epsilon_i$ with zero mean and variance $\sigma^2$ is added independently to each of the input variables $x_i$. By making use of $E[\epsilon_i] = 0$ and $E[\epsilon_i \epsilon_j] = \delta_{ij}\sigma^2$ (where $\delta_{ij} = 1$ when $i = j$ and 0 otherwise) show that minimizing $Err_D$ averaged over the noise distribution is equivalent to minimizing the sum-of-squares error for noise-free input variables with the addition of a weight-decay regularization term, in which the bias parameter $w_0$ is omitted from the regularizer.

(b) (HTF[1] Ex. 3.12 in on-line version, with some notation changed)

Show that the ridge regression estimates can be obtained by ordinary least squares regression on an augmented data set. First we *center* the data, computing a new matrix C, where $c_j^{(i)} = (x_j^{(i)} - \bar{x}_j)$; that is, that we subtract the average value of each feature j, $\bar{x}_j$, from each of the values for feature j in the data.

We augment the centered matrix C with d additional rows $\sqrt{\lambda}I$, and augment $\mathbf{y}$ with d zeros. By introducing artificial data having response value zero, the fitting procedure is forced to shrink the coefficients toward zero. This is related to the idea of *hints* due to Abu-Mostafa (1995), where model constraints are implemented by adding artificial data examples that satisfy them.

(c) (based on HTF Ex. 3.6 in on-line version, with some notation changed)
Show that the ridge regression estimate is the mean (and mode) of the posterior distribution, under a Gaussian prior $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \tau I)$, and Gaussian sampling model $\mathbf{y} \sim \mathcal{N}(X\mathbf{w}, \sigma^2 I)$. Find the relationship between the regularization parameter $\lambda$ in the ridge formula, and the variances $\tau$ and $\sigma^2$. Assume that the data are "centered" as described in the previous problem, so that we don't need a bias term.

[1] Hastie, Tibshirani and Friedman, The Elements of Statistical Learning (http://statweb.stanford.edu/ tibs/ElemStatLearn/)