Statistical Machine Learning Notes 09

Probabilistic Learning

Instructor: Justin Domke

Contents

1 Introduction **2**

2 Maximum Likelihood **2**

3 Examples of Maximum Likelihood **3**

3.1 Binomial . . . 3

3.2 Uniform Distribution . . . 4

3.3 Univariate Gaussian . . . 4

3.4 Multivariate Gaussian . . . 6

3.5 Spherical Multivariate Gaussian . . . 7

4 Properties of Maximum Likelihood **8**

4.1 Maximum Likelihood is Consistent . . . 8

4.2 Maximum Likelihood is Equivariant . . . 8

4.3 Maximum Likelihood is Efficient . . . 9

4.3.1 Proof of the Cramer-Rao bound . . . 10

4.4 Maximum Likelihood Assumes a Whole Lot . . . 11

4.5 Maximum Likelihood is Empirical Risk Minimization of the KL-divergence . . . 12

5 Bayesian Methods **13**

1 Introduction

Almost all of our methods for learning have been based off the function of "risk" and "loss". We have worked by picking some class of functions $f(\mathbf{x})$ mapping from inputs to outputs. We quantified how we wanted that function to behave in terms of the true risk

$$R_{\text{true}}(f) = \mathop{\mathbb{E}}_{p_0}\big[L(f(\mathbf{x}), \mathbf{y})\big] = \int\int p_0(\mathbf{x}, \mathbf{y}) L(f(\mathbf{x}), \mathbf{y}) d\mathbf{x} d\mathbf{y}, \tag{1.1}$$

where p_0 is the true (unknown) distribution. Then, we approximated this by an empirical risk, fit the function f, and we were done. Gazing at Eq. 1.1, however, another possible strategy comes to mind. Namely, why don't we approximate p_0 with some function p? Then, when we need to make predictions for some specific input $\mathbf{x}$, we can pick the best guess y' by doing

$$\min_{\mathbf{y}'} \int p(\mathbf{x}, \mathbf{y}) L(\mathbf{y}', \mathbf{y}) d\mathbf{y}.$$

Then, assuming that we have approximated p_0 perfectly, we could do everything exactly!

The fundamental question here is: when should we apply the loss function? In the traditional strategy, we apply it at training time: the predictor $f(\mathbf{x})$ is fit to give the best possible performance, with the loss "baked in". Now, we apply the loss function only at test time. Notice that we could even change loss functions "on the fly". We could also flip things around. If we suddenly decided we would rather predict $\mathbf{x}$ from $\mathbf{y}$, we could also do that.

This may seem very attractive. As we will see, however, there is a price to be paid for this generality. However, we postpone discussion of all the tradeoffs until later. The immediate question is more basic: how should we fit p?

(**Note**: when fitting $p(\mathbf{x}, \mathbf{y})$, notice that $\mathbf{x}$ and $\mathbf{y}$ are on an even footing. Thus, for simplicity, we will usually write the variables together as a single vector $\mathbf{x}$.)

2 Maximum Likelihood

There have been many methods proposed to fit distributions. In this class, we will focus on the "maximum likelihood" method. Suppose that we are fitting a distribution $p(\mathbf{x}; \boldsymbol{\theta})$, parametrized by some vector $\boldsymbol{\theta}$. Where does this distribution p come from? You pick it. How do you pick it? We will come back to that! Let the data be a set of vectors $\{\hat{\mathbf{x}}\}$.

The **log-likelihood** is

$$l(\boldsymbol{\theta}) = \hat{\mathbb{E}} \log p(X; \boldsymbol{\theta}) = \sum_{\hat{\mathbf{x}}} \log p(\hat{\mathbf{x}}; \boldsymbol{\theta}).$$

The **maximum likelihood** method, surprisingly enough, consists of picking $\boldsymbol{\theta}$ to maximize the likelihood

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}).$$

This method has some nice properties, but before worrying about them, let's try some examples. You may have seen these before in a statistics class.

3 Examples of Maximum Likelihood

3.1 Binomial

A binomial distribution is a distribution over a binary variable, with $x \in \{0, 1\}$, given by

$$p(x; \theta) = \theta^x (1 - \theta)^{1-x}.$$

Given some training data, we can calculate

$$l(\theta) = \sum_{\hat{x}} \log p(\hat{x}; \theta) = \sum_{\hat{x}} \big(\hat{x} \log \theta + (1 - \hat{x}) \log(1 - \theta) \big)$$

Now, we can maximize this by setting the derivative with respect to θ to zero. We have

$$\begin{aligned} \frac{dl}{d\theta} = 0 &= \sum_{\hat{x}} \big(\hat{x} \frac{1}{\theta} - (1 - \hat{x}) \frac{1}{1 - \theta} \big) \\ &= \#[\hat{x} = 1] \frac{1}{\theta} - \#[\hat{x} = 0] \frac{1}{1 - \theta}, \end{aligned}$$

where $\#[\hat{x} = 1]$ is the number of points in the training data with $\hat{x} = 1$. This equation is solved by

$$\theta = \frac{\#[\hat{x} = 1]}{\#[\hat{x} = 1] + \#[\hat{x} = 0]}.$$

Thus, the maximum likelihood estimate is that the binomial distribution has the same probability of being 1 as in the training data. This is quite intuitive.

3.2 Uniform Distribution

Consider the distribution uniform on 0 to θ.

$$p(x;\theta) = \frac{1}{\theta} I[0 \leq x \leq \theta]$$

The log likelihood is, as ever,

$$l(\theta) = \sum_{\hat{x}} \log p(\hat{x}, \theta).$$

If θ is less than any value $\hat{x}$, then the probability of that point is zero, and we can think of the log-likelihood as being $-\infty$. On the other hand, suppose that

$$\theta > \hat{x}, \ \forall \hat{x}.$$

Then, we have that

$$l(\theta) = \sum_{\hat{x}} \log \frac{1}{\theta} = -\sum_{\hat{x}} \log \theta.$$

Thus, the likelihood is maximized by setting

$$\theta = \max_{\hat{x}} \hat{x}.$$

3.3 Univariate Gaussian

A univariate gaussian distribution is defined by

$$p(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\big(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\big).$$

Through a bunch of manipulation, we can take the logarithm of this, and then the derivatives of the logarithm.

First of all, we will use the fact that

$$\log \sqrt{2\pi\sigma^2} = \frac{1}{2} \log(2\pi\sigma^2).$$

Then, taking the logarithm and then derivatives, we have

$$\begin{aligned}
\log p(x;\mu,\sigma^2) &= -\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2} - \frac{1}{2}\log\sigma^2 - \frac{1}{2}\log 2\pi \\
\frac{d}{d\mu}\log p(x;\mu,\sigma^2) &= \frac{(x-\mu)}{\sigma^2} \\
\frac{d}{d\sigma^2}\log p(x;\mu,\sigma) &= \frac{1}{2}\frac{(x-\mu)^2}{(\sigma^2)^2} - \frac{1}{2}\frac{1}{\sigma^2}.
\end{aligned}$$

Now, we want to do maximum likelihood estimation. That is, we need to maximize

$$\max_{\mu,\sigma^2} l(\mu,\sigma^2)$$

We can do this by solving the two equations

$$\begin{aligned}
\frac{d}{d\mu}l(\mu,\sigma^2) = &\quad \textstyle\sum_{\hat{x}} \frac{d}{d\mu}\log p(\hat{x};\mu,\sigma) &= 0 \\
\frac{d}{d\sigma^2}l(\mu,\sigma^2) = &\quad \textstyle\sum_{\hat{x}} \frac{d}{d\sigma^2}\log p(\hat{x};\mu,\sigma) &= 0
\end{aligned}$$

From the first condition, it is easy to see that

$$\mu = \mathrm{mean}_{\hat{x}}\hat{x}$$

From the second condition, we can then find

$$\begin{aligned}
0 &= \sum_{\hat{x}}\Big(\frac{1}{2}\frac{(\hat{x}-\mu)^2}{(\sigma^2)^2} - \frac{1}{2}\frac{1}{\sigma^2}\Big) \\
&= \sum_{\hat{x}}\Big(\frac{(\hat{x}-\mu)^2}{\sigma^2} - 1\Big) \\
&= \sum_{\hat{x}}\big((\hat{x}-\mu)^2 - \sigma^2\big) \\
\sigma^2 &= \mathrm{mean}_{\hat{x}}(\hat{x}-\mu)^2
\end{aligned}$$

3.4 Multivariate Gaussian

A multivariate Gaussian is defined by

$$\begin{aligned} p(\mathbf{x};\boldsymbol{\mu},\Sigma) &= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\bigl(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\bigr) \\ \log p(\mathbf{x};\boldsymbol{\mu},\Sigma) &= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) - \frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma^{-1}|. \end{aligned}$$

It is an unfortunate, but firmly established convention to use the symbol "Σ" to denote the covariance matrix. It is important not to get this confused with a sum. (In these notes, the difference is indicated by the size of the symbol, as well as context.)

First off, let's calculate some properties of this distribution. It is not hard to see that, by symmetry,

$$\mathop{\mathbb{E}}_{p}[\mathbf{x}] = \boldsymbol{\mu}.$$

It can also be shown that

$$\mathop{\mathbb{E}}_{p}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T] = \Sigma.$$

Thus, it makes sense to call $\boldsymbol{\mu}$ the "mean" and Σ the "covariance matrix". In order to calculate the maximum likelihood estimate, we will need some derivatives. Using the fact that Σ^{-1} is symmetric,

$$\frac{d}{d\boldsymbol{\mu}}\log p(\mathbf{x};\boldsymbol{\mu},\Sigma) = \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

Using the fact that $\frac{d}{dX}\mathbf{a}^T X\mathbf{a} = \mathbf{a}\mathbf{a}^T$ and the strange but true fact that $\frac{d\log|X|}{dX} = X^{-T}$, and again assuming that Σ is symmetric, we have

$$\frac{d}{d\Sigma^{-1}}\log p(\mathbf{x};\boldsymbol{\mu},\Sigma) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T + \frac{1}{2}\Sigma.$$

Now, as ever, when doing maximum likelihood estimation, our goal is to accomplish the maximization

$$\max_{\Sigma,\boldsymbol{\mu}} l(\Sigma, \boldsymbol{\mu}) = \max_{\Sigma,\boldsymbol{\mu}} \sum_{\hat{\mathbf{x}}} \log p(\hat{\mathbf{x}}; \boldsymbol{\mu}, \Sigma).$$

Setting $dl/d\boldsymbol{\mu} = \mathbf{0}$, we have

$$\begin{aligned} \sum_{\hat{\mathbf{x}}} \frac{d}{d\boldsymbol{\mu}} \log p(\hat{\mathbf{x}}; \boldsymbol{\mu}, \Sigma) &= \sum_{\hat{\mathbf{x}}} \Sigma^{-1}(\hat{\mathbf{x}} - \boldsymbol{\mu}) = \mathbf{0} \\ \boldsymbol{\mu} &= \boxed{\text{mean}_{\hat{\mathbf{x}}} \hat{\mathbf{x}}.} \end{aligned}$$

Setting $dl/d\Sigma^{-1} = 0$, we have

$$\begin{aligned} \sum_{\hat{\mathbf{x}}} \frac{d}{d\Sigma^{-1}} \log p(\hat{\mathbf{x}}; \boldsymbol{\mu}, \Sigma) &= \sum_{\hat{\mathbf{x}}} \big(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T + \frac{1}{2}\Sigma\big) \\ \Sigma &= \boxed{\text{mean}_{\hat{\mathbf{x}}} (\hat{\mathbf{x}} - \boldsymbol{\mu})(\hat{\mathbf{x}} - \boldsymbol{\mu})^T.} \end{aligned}$$

Again, this is all very intuitive. The mean is the empirical mean, and the covariance matrix is the empirical covariance matrix. However, this is *not* unbiased. (Recall to estimate the variance of a scalar variable, we should use the formula $\frac{1}{n-1}\sum_{i=1}^n (x_i - \mu)^2$, rather than the empirical variance $\frac{1}{n}\sum_{i=1}^n (x_i - \mu)^2$.) So maximum likelihood will tend to slightly overestimate the variance when the number of data is small.

3.5 Spherical Multivariate Gaussian

A spherical Gaussian is just a Gaussian distribution where we constrain the covariance matrix to take the form

$$\Sigma = aI$$

for some constant a. Using the fact that $|aI| = a^d$, this is

$$p(\mathbf{x}; \boldsymbol{\mu}, a) = \frac{1}{(2\pi)^{d/2} a^{d/2}} \exp\big(-\frac{1}{2a}(\mathbf{x} - \boldsymbol{\mu})^T (\mathbf{x} - \boldsymbol{\mu})\big).$$

This turns out to have a maximum likelihood solution of

$$\boldsymbol{\mu} = \text{mean}_{\hat{\mathbf{x}}} \hat{\mathbf{x}}.$$

$$a = \frac{1}{d}\text{mean}_{\hat{\mathbf{x}}}(\hat{\mathbf{x}} - \boldsymbol{\mu})^T(\hat{\mathbf{x}} - \boldsymbol{\mu}).$$

4 Properties of Maximum Likelihood

Here we will informally discuss some of the properties of maximum likelihood.

4.1 Maximum Likelihood is Consistent

If the data is actually being generated by a distribution

$$p(\mathbf{x}; \boldsymbol{\theta}_0),$$

for some vector $\boldsymbol{\theta}_0$, then (absent pathological conditions) as the amount of data goes to infinity, the parameters $\boldsymbol{\theta}$ recovered by maximum likelihood will converge to $\boldsymbol{\theta}_0$. This is a definitely a good property, as we probably would consider any method lacking it to be, more or less, broken.

4.2 Maximum Likelihood is Equivariant

Another nice property of the likelihood is that it is "equivariant". This just means that we can reparameterize without affecting the solution. Specifically, suppose we are considering estimating some distribution

$$p(\mathbf{x}; \boldsymbol{\theta}).$$

Suppose the maximum likelihood estimate of $\boldsymbol{\theta}$ on some dataset is $\boldsymbol{\theta}^*$. Now, we choose to instead parametrize our function by $\boldsymbol{\phi}$, which is some nonlinear transformation of $\boldsymbol{\theta}$.

$$\boldsymbol{\theta} = \mathbf{g}(\boldsymbol{\phi})$$

Now, if we define

$$q(\mathbf{x}; \boldsymbol{\phi}) = p\big(\mathbf{x}; \mathbf{g}(\boldsymbol{\phi})\big)$$

and do maximum likelihood estimation of $\boldsymbol{\phi}$, we will recover $\boldsymbol{\phi}^*$ such that

$$\boldsymbol{\theta}^* = \mathbf{g}(\boldsymbol{\phi}^*).$$

(Proving this is quite easy.) Again, this is a reassuring property: The exact details of how we have parametrized our function don't matter. Failing to be equivariant wouldn't seem to be quite so disqualifying as failing to be consistent, but it is certainly comforting.

4.3 Maximum Likelihood is Efficient

Perhaps the strongest argument in favor of maximum likelihood is that it is asymptotically *efficient*. Suppose the data is actually being generated by a distribution

$$p(\mathbf{x}; \boldsymbol{\theta}_0),$$

for some vector $\boldsymbol{\theta}_0$. As discussed above, the maximum likelihood is consistent, in the sense that it converges to $\boldsymbol{\theta}_0$. The next question is, *how fast* does it do that? Is there some other measure that converges faster?

Asymptotically, the answer is no. This result hinges on defining "faster" as the expected squared distance between our estimate $\boldsymbol{\theta}$ and the true parameters $\boldsymbol{\theta}_0$. This follows from two results that are described here informally. (For simplicity, these are stated here for a scalar parameter θ.) These make use of a quantity called the Fisher information.

$$I(\theta_0) = \mathop{\mathbb{E}}_{p(X;\theta_0)}\Big[\Big(\frac{\partial \log p(X;\theta_0)}{\partial \theta}\Big)^2\Big]$$

Intuitively, we can understand this. Consider a "landscape" of different values θ, in which we seek to locate the true value θ_0. If log-likelihood changes a lot in the region around θ_0, then we should expect the true parameters to be relatively easy to locate.

The two results showing the efficiency of maximum likelihood are:

1. The Cramer-Rao bound. This states that no unbiased estimator can have a variance less than $\frac{1}{nI(\theta_0)}$. (Technically, maximum likelihood is not unbiased, but this is good enough for our purposes, since we are looking for an asymptotic result anyway.)
2. The asymptotic normality the maximum likelihood. This shows that, as the amount of data becomes large, the estimated parameters will be distributed with variance $\frac{1}{nI(\theta_0)}$. Specifically, they will be distributed as a Gaussian distribution with this variance centered at θ_0.

4.3.1 Proof of the Cramer-Rao bound

(This sub-sub-section may be considered recreational.) Consider any unbiased estimator $\hat{\theta}(X_1, ..., X_n)$. By definition,

$$\mathop{\mathbb{E}}_{p(X_1;\theta)} \mathop{\mathbb{E}}_{p(X_2;\theta)} ... \mathop{\mathbb{E}}_{p(X_n;\theta)} (\hat{\theta}(X_1, ..., X_n) - \theta) = 0.$$

Written in integral form, this is

$$\int_{x_1} ... \int_{x_n} p(x_1;\theta)...p(x_n;\theta) \, (\hat{\theta}(x_1, ..., x_n) - \theta) = 0.$$

If we differentiate this with respect to θ, noting that $\frac{d \log p(x_i;\theta)}{d\theta} = \frac{1}{p(x_i;\theta)} \frac{dp(x_i;\theta)}{d\theta}$, we have

$$\int_{x_1} ... \int_{x_n} p(x_1;\theta)...p(x_n;\theta) \Big(\sum_i \frac{d \log p(x_i;\theta)}{d\theta} \Big) \, (\hat{\theta}(X_1, ..., X_n) - \theta) = 1,$$

which, when squared, is

$$\Big(\int_{x_1} ... \int_{x_n} p(x_1;\theta)...p(x_n;\theta) \Big(\sum_i \frac{d \log p(x_i;\theta)}{d\theta} \Big) \, (\hat{\theta}(X_1, ..., X_n) - \theta) \Big)^2 = 1.$$

Now, we will define inner products as expectations. Let f and g be functions of $x_1, ..., x_n$, and let the inner product between them be defined by taking an expectation over p of their product, namely

$$f \cdot g = \int_{x_1} ... \int_{x_n} p(x_1;\theta)...p(x_n;\theta) f(x_1, ..., x_n) g(x_1, ..., x_n).$$

(It is easy to verify that this obeys the necessary notions of an inner product, i.e. that $f \cdot g = g \cdot f$, $(cf) \cdot g = c(f \cdot g)$, $(f + g) \cdot h = f \cdot h + g \cdot h$, and $f \cdot f \geq 0$, where f, g, and h are functions, and c is a real number.)

The Cauchy-Scharz inequality states that $(f \cdot g)^2 \leq (f \cdot f)(g \cdot g)$. Applying this to our above equation gives us

$$\Big(\int_{x_1} ... \int_{x_n} p(x_1;\theta)...p(x_n;\theta) \Big(\sum_i \frac{d \log p(x_i;\theta)}{d\theta} \Big)^2 \Big) \times \Big(\int_{x_1} ... \int_{x_n} p(x_1;\theta)...p(x_n;\theta) (\hat{\theta}(X_1, ..., X_n) - \theta)^2 \Big) \geq 1. \quad (4.1)$$

Now, attack the first part of the equation. This is

$$\begin{aligned}
&\int_{x_1} ... \int_{x_n} p(x_1;\theta)...p(x_n;\theta) \sum_i \sum_j \frac{d \log p(x_i;\theta)}{d\theta} \frac{d \log p(x_j;\theta)}{d\theta} \\
= \quad &\sum_i \int_{x_i} p(x_i;\theta) \big(\frac{d \log p(x_i;\theta)}{d\theta}\big)^2 + \sum_i \int_{x_i} p(x_i;\theta) \frac{d \log p(x_i;\theta)}{d\theta} \sum_{j \neq i} \int_{x_j} p(x_j;\theta) \frac{d \log p(x_j;\theta)}{d\theta}.
\end{aligned}$$

However, we know that,

$$\int_x p(x;\theta) \frac{d \log p(x;\theta)}{d\theta} = \int_x p(x;\theta) \frac{1}{p(x;\theta)} \frac{dp(x;\theta)}{d\theta} = \frac{d}{d\theta} \int_x p(x;\theta) = 0,$$

and so, we have

$$\int_{x_1} ... \int_{x_n} p(x_1;\theta)...p(x_n;\theta) \big(\sum_i \frac{d \log p(x_i;\theta)}{d\theta}\big)^2 = np(x;\theta) \big(\frac{d \log p(x_i;\theta)}{d\theta}\big)^2 = nI(\theta).$$

Next, it is easy to see that the second part of the left-hand side of Eq. 4.1 is equal to $\mathbb{E}(\hat{\theta} - \theta)^2$. Thus, we have the Cramer-Rao bound,

$$\boxed{\mathop{\mathbb{E}}_{p(X_1;\theta)} ... \mathop{\mathbb{E}}_{p(X_n;\theta)} (\hat{\theta}(X_1, ..., X_n) - \theta)^2 \geq \frac{1}{nI(\theta)}}.$$

Note: the Fisher information is sometimes defined in slightly different ways, in which the n in the above equation would be absorbed into the information I.

4.4 Maximum Likelihood Assumes a Whole Lot

Now, suppose the true distribution is

$$p_0(\mathbf{x}).$$

Most of the above properties have hinged on the assumption that there exists a vector $\boldsymbol{\theta}_0$ such that

$$p_0(\mathbf{x}) = p(\mathbf{x}; \boldsymbol{\theta}_0).$$

Another way of stating this is that we have a **well-specified** model.

You might ask: how could we ever know this? The brief answer is that we probably don't. Now, we can create somewhat contrived situations where it is true. It is hard to see how a binary variable can fail to be Binomial! In general, however, making a model tends to be an educated guess of sorts.

4.5 Maximum Likelihood is Empirical Risk Minimization of the KL-divergence

All of the discussion has depended on the assumption that the true data-generating distribution is known. As this is almost never true is practice, it might seem like maximum likelihood could almost never be used!

Unfortunately, without the assumption of a well-specified model, almost of the above properties disappear. After all, how can $\boldsymbol{\theta}$ converge to $\boldsymbol{\theta}_0$ if $\boldsymbol{\theta}_0$ doesn't exist?

On the other hand, intuitively, it seems like the maximum likelihood should still do something reasonable if the model is "almost" well-specified. That is, if there exists some vector of parameters $\tilde{\boldsymbol{\theta}}$ such that $p_0(\mathbf{x}) \approx p(\mathbf{x}; \tilde{\boldsymbol{\theta}})$, shouldn't maximum likelihood converge to something close to p_0? After all, in the previous case, the data *could have* come from parameters $\tilde{\boldsymbol{\theta}}$– how could the maximum likelihood even "know" it did not?

In fact, maximum likelihood does behave reasonably in the face of minor misspecification. To understand this, we must first introduce the The **Kullback-Leibler divergence**.

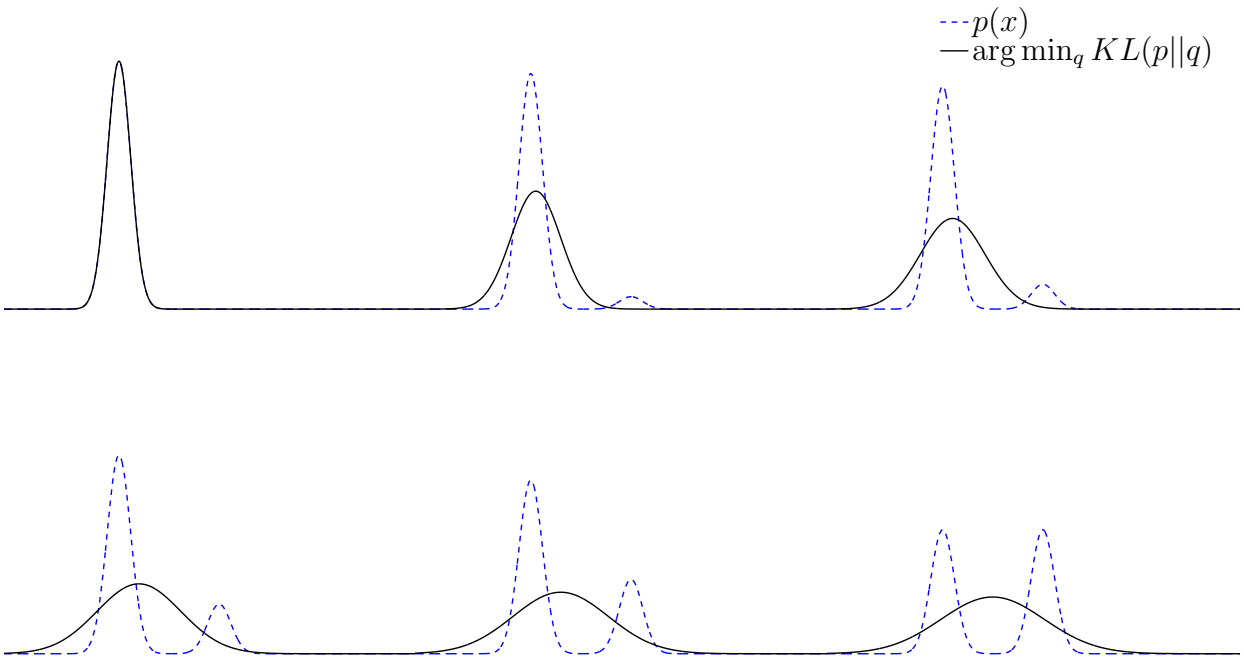
$$KL(p_0||p) = \int_{\mathbf{x}} p_0(\mathbf{x}) \log \frac{p_0(\mathbf{x})}{p(\mathbf{x})}$$

This is a sort of "divergence measure between probability distributions". Its origins come from information theory[1]. The important thing to note about the KL-divergence is that it is non-negative, and zero only when $p_0 = p$. Note also that it is not actually a distance measure, as it is not symmetric.

Suppose that there is some region of points where p_0 is significant, but p is near zero. As the KL-divergence measures the logarithm of p, this region leads to a large divergence.

The following figures show several base distributions p_0 (shown as dotted lines). For each distribution, the Gaussian is computed that minimizes the KL-divergence (shown as solid lines).

[1] Where it can be thought of as measuring "the expected number of bits wasted if you build a code for $\mathbf{x}$ assuming that the distribution is p when the actual distribution is p_0"



Now, consider the KL-divergence between the true distribution p_0, and the one that we fit, p.

$$\begin{aligned}
\arg\min_p KL(p_0||p) &= \arg\min_p \sum_{\mathbf{x}} p_0(\mathbf{x}) \log p_0(\mathbf{x}) - \sum_{\mathbf{x}} p_0(\mathbf{x}) \log p(\mathbf{x}) \\
&= \arg\max_p \sum_{\mathbf{x}} p_0(\mathbf{x}) \log p(\mathbf{x}) \\
&\approx \arg\max_p \sum_{\hat{x}} \log p(\hat{x})
\end{aligned}$$

In third line we have made essentially an empirical approximation of the true risk above.

The way to understand this is that if the true distribution is any of the dotted curves above and we fit a Gaussian then, as the amount of data increases, we will recover the solid curve.

5 Bayesian Methods

Suppose we have a big jar full of bent coins. We happen to know that, inside of this bin, there are 75 coins of type A with probability 60% and 25 coins of type B that come up with probability 40%. Now, we pick a coin at random out of the jar. We flip it 8 times, and observe 3 heads, followed by 5 tails. What is the probability that we have in our hands coin of type A?

One approach to this is to apply Bayes' theorem.

$$Pr(X|Y) = \frac{Pr(Y|X)Pr(X)}{Pr(Y)}$$

In our case, we want to calculate the probability that we have a coin of type A, given that we have observed 3 heads in 10 coin flips.

$$Pr(A|\text{Data}) = \frac{Pr(\text{Data}|A)Pr(A)}{Pr(\text{Data})}$$

$$Pr(B|\text{Data}) = \frac{Pr(\text{Data}|B)Pr(B)}{Pr(\text{Data})}$$

Now, in our case, we know that we have a 75% chance of grabbing a coin of type A.

$$Pr(A) = .75$$

$$Pr(B) = .25$$

Now, if we had a coin of type A, we would have probability

$$Pr(\text{Data}|A) = .6^3 \cdot .4^5 \approx 0.00221$$

$$Pr(\text{Data}|B) = .4^3 \cdot .6^5 \approx 0.00498$$

Thus, we have

$$Pr(A|\text{Data}) = 0.00165/Pr(\text{Data})$$

$$Pr(B|\text{Data}) = 0.00124/Pr(\text{Data}).$$

Now, notice that we don't need to go through too much calculation to recover $Pr(\text{Data})$. Since we know that

$$Pr(A|\text{Data}) + Pr(B|\text{Data}) = 1,$$

we can just normalize and calculate

$$P(A|\text{Data}) = .57$$

$$P(B|\text{Data}) = .43.$$

Thus, there is a 57% chance we have a coin of type A.

Now, all of the above may seem quite uncontroversial. However, when we say there is a 57% chance our coin is of type A, what exactly does that mean? After all, we picked one *particular* coin. It is either of type A or it isn't. What probabilities exactly are we talking about here? The traditional view holds that talking about such probabilities is meaningless. On the other hand, forced to bet, wouldn't everyone choose A[2]? There are philosophical issues here about the meaning of probability. We won't get too deeply into these, however, just note that they exist, and are a part of the debate in statistics between Bayesian and traditional frequentist methods.

Now, let's try to formalize the process that we used above and scale it up to larger problems. Instead of just two types of coins (two different binomial distributions), imagine we have a set of potential probability distributions p. Imagine also that we have, by some prior knowledge, a distribution $Pr(p)$ over these distributions. What happens is the following:

- Some distribution p is picked, with probability proportional to $Pr(p)$.
- A bunch of samples $\{\hat{\mathbf{x}}\}$ is drawn from p.
- We get to see $\{\hat{\mathbf{x}}\}$, and need to make predictions about the future.

The simplest way to approach this situation is to again apply Bayes' equation

$$Pr(p|\{\hat{\mathbf{x}}\}) = \frac{Pr(\{\hat{\mathbf{x}}\}|p)Pr(p)}{Pr(\{\hat{\mathbf{x}}\})}.$$

Now, it makes sense to try to recover the most probable p. This means searching for

$$\begin{aligned}
\arg\max_p Pr(p|\{\hat{\mathbf{x}}\}) &= \arg\max_p Pr(\{\hat{\mathbf{x}}\}|p)Pr(p) \\
&= \arg\max_p \log Pr(\{\hat{\mathbf{x}}\}|p) + \log Pr(p) \\
&= \arg\max_p \log \prod_{\hat{\mathbf{x}}} p(\hat{\mathbf{x}}) + \log Pr(p) \\
&= \arg\max_p \sum_{\hat{\mathbf{x}}} \log p(\hat{\mathbf{x}}) + \log Pr(p).
\end{aligned}$$

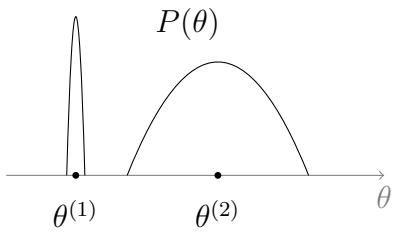
[2]Is it a contradiction to choose A and yet reject the idea of probabilities like this?

(In the first line, we exploit the fact that $Pr(\{\hat{\mathbf{x}}\})$ is constant with respect to p and so does not affect the maximizing p. In the second line, we take the logarithm. In the third line, we use the fact that $Pr(\{\hat{\mathbf{x}}\}|p) = \prod_{\hat{\mathbf{x}}} Pr(\hat{\mathbf{x}}|p) = \prod_{\hat{\mathbf{x}}} p(\hat{\mathbf{x}})$. The fourth line is just algebra.)

Thus, in the last line, we just have the log-likelihood plus the log priod $Pr(p)$. Searching for p to maximize $Pr(p|\{\hat{\mathbf{x}}\})$ is known as **maximum a posteriori** (MAP) estimation.

Notice the similarity to regularized maximum likelihood estimation. For example, it is common to parameterize p by some vector $\boldsymbol{\theta}$, and set $Pr(p) = Pr(\boldsymbol{\theta})$ to be a Gaussian centered at the origin. It is easy to show that doing this results in $\log Pr(\boldsymbol{\theta}) = -a||\boldsymbol{\theta}||^2$, where a depends on the covariance of the Gaussian. Similarly, it can be shown that the lasso penalty corresponds to a distribution of the form $Pr(\boldsymbol{\theta}) \propto \exp\big(-a||\boldsymbol{\theta}||_1\big)$. Thus, many Bayesians view regularized maximum likelihood estimation as "implicit" MAP estimation.

We should note, though, that "real" Bayesians *do not do MAP estimation*. To understand why not, suppose that we have a probability distribution parameterized by a scalar, and the postereior $Pr(\theta|\{\hat{x}\})$ looks something like the following:



MAP estimation will choose $\theta^{(1)}$ as the most probable set of parameters. However, this doesn't look so good, since most of the probability is in the area of $\theta^{(2)}$.

What real Bayesians do is not estimate one particular distribution p, but rather, make predictions *directly from the posterior* $Pr(p|\text{Data})$.

How is this done? Let's look at an example. Suppose we need to guess one single value for $\mathbf{x}$. Consider the loss of some guess $\mathbf{x}'$:

$$\min_{\mathbf{x}'} \int_{\mathbf{x}} L(\mathbf{x}', \mathbf{x}) Pr(\mathbf{x}|\text{Data}) d\mathbf{x}.$$

Now, we can calculate the probability of some particular output $\mathbf{x}$ by integrating over the possible

$$\begin{aligned} Pr(\mathbf{x}|\text{Data}) &= \int_{\boldsymbol{\theta}} Pr(\mathbf{x}, \boldsymbol{\theta}|\text{Data}) d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} Pr(\mathbf{x}|\boldsymbol{\theta}) Pr(\boldsymbol{\theta}|\text{Data}) d\boldsymbol{\theta} \\ &\propto \int_{\boldsymbol{\theta}} Pr(\mathbf{x}|\boldsymbol{\theta}) Pr(\text{Data}|\boldsymbol{\theta}) Pr(\boldsymbol{\theta}) d\boldsymbol{\theta}. \end{aligned}$$

Thus, finally, the true Bayesian chooses their best guess $\mathbf{x}'$ by the problem

$$\min_{\mathbf{x}'} \int_{\mathbf{x}} \int_{\boldsymbol{\theta}} L(\mathbf{x}', \mathbf{x}) Pr(\mathbf{x}|\boldsymbol{\theta}) Pr(\text{Data}|\boldsymbol{\theta}) Pr(\boldsymbol{\theta}) d\mathbf{x} d\boldsymbol{\theta}. \tag{5.1}$$

The question is, how to do this integral? In some situations, this can be done in closed form. In general, however, one must resort to Markov chain Monte Carlo techniques for approximately doing the integral[3]. This can be quite computationally challenging, which can be a major drawback of Bayesian methods.

Let's consider the advantages and disadvantages of the Bayesian approach.

- The major advantage of this is that it is, in a certain sense, the optimal method. If the true distribution is drawn from $Pr(\boldsymbol{\theta})$ then, on average, no method for making predictions can have lower loss than Eq. 5.1. To make this precise, suppose that we repeatedly get parameters $\boldsymbol{\theta}$ from the distribution $Pr(\boldsymbol{\theta})$, sample some data from $Pr(\text{Data}|\boldsymbol{\theta})$, make a predicted $\mathbf{x}'$, then measure the loss $L(\mathbf{x}', \mathbf{x})$ on some new $\mathbf{x}$ drawn from $p(\mathbf{x}, \boldsymbol{\theta})$. The above recipe will have the lowest average loss of any method. For this reason, many people feel that Bayesian methods are the one, true way to do machine learning.

- A disadvantage of Bayesian methods is that they can often be quite computationally expensive. As mentioned above, in complex problems, it is common to use MCMC techniques to do inference. These techniques do have guarantees of eventually converging to the right answer, but these guarantees are usually asymptotic in nature. Thus, given a finite amount of running time, one can be unsure how close the current answer is to the best one. Research is ongoing on faster MCMC methods with an eye on Bayesian inference.

- The most obvious issue with Bayesian methods is the need to specify the prior $Pr(\boldsymbol{\theta})$. In real applications, where does this prior come from? This is similar to issue we faced when doing (non-Bayesian) probabilistic modeling– we needed to specify a correct parametric model $p(\mathbf{x}; \boldsymbol{\theta})$. While specifying the prior may appear to be a drawback of

[3]See "Introduction to Monte Carlo methods" by David MacKay for a good tutorial on these techniques.

Bayesian methods, it is also something of an advantage. If you have a lot of knowledge about a particular domain, and you are able to specify this knowledge as a prior, Bayesian methods provide a nice framework to combine your knowledge with knowledge gained from data. Note also that, in the view of some, techniques like regularization are essentially MAP estimation in all but name.

There is a great deal of material out there on the debate between frequentist and Bayesian statistics.