Unit 6: regularization, priors and Bayesian inference

Ethan Levien

October 29, 2025

1.1 Introduction to Bayesian inference: some simple examples

- Models encode our assumptions about how data was generate. Sometimes we want to incorporate vague information, such as "the function describing my data is very smooth and changes roughly on a time-scale of 5 hours". Or, we might want to include many predictors, but to avoid overfitting penalize high values of the predictors. For example, we might believe there is an interaction term, but suspect it is much smaller than the additive terms. This motivates a different approach to statistics, one which treats parameters as themselves random variables.

Example 1 (Bernoulli with priors)**.** Consider the model

$$X \sim \text{Bernoullli}(q)$$

An implicit assumption we have made when we fit this model, is that before we see any data, each value of q is equally likely. To make this more explicit, we can write

$$\begin{aligned} X|q &\sim \text{Bernoullli}(q) \\ q &\sim \text{Uniform}(0,1) \end{aligned}$$

Another way to think about constructing an estimator of q, is that given data points $X_1, \ldots, X_N$, we would like to know

$$\hat{q}_b = E[q|X_1, \ldots, X_N] \tag{1}$$

We can in principle solve this problem using Bayes' theorem. In fact, we would also take it a step further and ask: What is the distribution of

$$q|X_1, \ldots, X_N$$

Is turns out that this has a relatively simple form, called a β-distribution. We say

$$q \sim \text{Beta}(a, b)$$

if q is a random variable on $[0, 1]$ which has density

$$f(q) = Bq^{a-1}(1-q)^{b-1}$$

The constant B ensures the area under this curve between $q = 0$ and $q = 1$ is indeed one, as it must be for a random density.

Question: Show that

$$q|X_1, \ldots, X_N \sim \text{Beta}(Y+1, N-Y+1)$$

Solution: Using Bayes' theorem

$$f(q|Y) = \frac{P(Y|q)f(q)}{P(Y)}$$
$$= \frac{1}{P(Y)}\binom{N}{Y}q^Y(1-q)^{N-Y}$$

Since we are thinking of this as a probability density in q, we don't actually need to compute $P(Y)$ explicitly (even though we could), instead, we can simply notice that

$$f(q|Y) = Cq^Y(1-q)^{N-Y}$$

for some constant of proportionality C which will depend on Y, but is uniquely determined by the fact that the area under this curve must be 1. In the colab notebook we plot this curve.

- To summarize the idea of Bayesian inference, If D is a our data and θ is our parameter(s) Bayes' theorem tells us that

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \tag{2}$$

- The distributions appearing in Equation (2) have the following names and interpretations:
  - $P(\theta|D)$ is the posterior (a beta distribution in Example 1).
  - $P(D|\theta)$ is the likelihood (Binomial distribution in Example 1).
  - $P(\theta)$ is the prior distribution (Uniform distribution in Example 1)
  - $P(D)$ is the evidence. It represents the chance we observe the data *unconditional* on the parameters. This can be obtained by marginalizing over the priors.

  While in classical statistics, the objective is to determine (estimate) a parameter θ, in Bayesian statistics our objective is to compute the posteior distribution. Once we have this, we can obtain so-called point estimates, for example, by taking the average of θ or maximum of $P(\theta|D)$ (maximum likelihood).

- The beta distribution will play an important role in what follows, so we note that if

$$q \sim \text{Beta}(a, b)$$

  then (using calculus) it can be shown that

$$E[q] = \frac{a}{a+b}$$
$$\text{var}(q) = \frac{ab}{(a+b)^2(a+b+1)}.$$

  Notice that this implies the estimator $\hat{q}_b$ in Equation 1 becomes

$$\hat{q}_b = \frac{Y+1}{Y+1+N-Y+1} = \frac{Y+1}{N+2}$$

  This is in-fact the estimator $\hat{q}_L$ from the week 7 exercise set!

- The mode – the value which maximizes the probability density – for the β-distribution is

$$\frac{a-1}{a+b-2}$$

  We can see that the estimator $\hat{q} = Y/N$ comes comes from the mode $f(q|Y)$ – recall that such an estimator is called the maximum likelihood estimator.

- A very important observation, which is what the problem on the week 7 exercise set is all about, is that adding priors often has the effect of introducing a bias while reducing the variance.
- Another important feature of the β distribution is that when a and b are very large, it is approximately Normal with this mean and variance. This means we can get a rough idea of the probabilities for a β random variable. Equipped with this knowledge, we can try to go beyond the case where our prior assumption about the distribution of q, before fitting the model, is Uniform.

Example 2 (Selecting priors)**.** Suppose we are given a coin. We are trying to determine whether it is biased based on the outcome of N flips. We are pretty sure that, like most coins, it is not biased. To be precise, let's say you are 95% confident that the bias of the coin is less than 10% biased towards heads or tails.

Question:

(a) Select a prior distribution for q.

(b) What is the posterior expectation if we flip the coin 5 times and see 5 heads?

Solution:

(a) We know that a β-distribution is approximately Normal if a and b are large enough. Assuming we can make a Normal approximation, we can use the formulas for the mean and variance of the b distribution to select parameters such that

$$P(|q - 0.5| < 0.1) = P(0.4 < q < 0.6) \approx 0.95 \tag{3}$$

Since we would like the mean of our prior distribution to be $1/2$, we select

$$q \sim \text{Beta}(a, a).$$

If q is approximately Normal, then Equation 3 will hold when

$$\begin{aligned} 1.96\sqrt{\text{var}(q)} &= 0.1 \\ &\implies \sqrt{\frac{a^2}{4a^2(2a+1)}} = \frac{1}{2}\sqrt{\frac{1}{2a+1}} = \frac{0.1}{1.6} \\ &\implies a \approx 31.5 \end{aligned}$$

(b) In order to compute the posterior, we use Bayes' theorem

$$f(q|X_1, \ldots, X_N) \propto q^Y(1-q)^{N-Y} \times q^{a-1}(1-q)^{a-1} = q^{Y+a-1}(1-q)^{N-Y+a-1}$$

where $Y = \sum_{i=1}^{N} X_i$. Again, the q dependence uniquely determines the distribution, since the area under this curve must be one. We can therefore say that

$$q|X_1, \ldots, X_N \sim \text{Beta}(Y + a, N - Y + a)$$

For the mean and variance we have

$$E[q|X_1, \ldots, X_N] = \frac{Y + a}{N + 2a} = \frac{5 + 31.5}{5 + 2 \times 31.5} \approx 0.54$$

Note that if we used usual estimator we would have found $\hat{q} = 1$, while if we used uniform priors and taken the posterior expectation $\hat{q}_L = 0.85$. Yet another approach would be to take the posterior maximum (the mode) of the posterior in using the β-distribution priors.

- Something quite nice about using β-distribution priors for the Bernoulli model is that BOTH the prior and the posterior have a β-distribution, albeit with difference parameters. In general, when this is the case we say that the priors are conjugate.

Example 3 (Bayesian inference with a Normal distribution)**.** Suppose we have a Normal model for Y

$$Y|(\mu,\sigma) \sim \text{Normal}(\mu,\sigma).$$

Assume that σ is known and take our priors on μ to be

$$\mu \sim \text{Normal}(\mu_0,\sigma_\mu)$$

Question: What is the posterior distribution of μ?

Solution: The likelihood is

$$\begin{aligned} p(D|\theta) &= \prod_i \frac{1}{\sqrt{2\sigma^2\pi}} e^{-(Y_i-\mu)^2/2\sigma^2} \\ &= \frac{1}{(2\sigma^2\pi)^{N/2}} e^{-\sum_i (Y_i-\mu)^2/2\sigma^2} \end{aligned}$$

The posterior is *proportional to*

$$\begin{aligned} p(D|\theta)p(\theta) &= \frac{1}{(2\sigma^2\pi)^{N/2}} e^{-\sum_i (Y_i-\mu)^2/2\sigma^2} \frac{1}{\sqrt{2\pi\sigma_\mu^2}} e^{-(\mu-\mu_0)^2/2\sigma_\mu^2} \\ &= \frac{1}{(2\sigma^2\pi)^{N/2}} \frac{1}{\sqrt{2\pi\sigma_\mu^2}} e^{-\sum_i (Y_i-\mu)^2/2\sigma^2 - (\mu-\mu_0)^2/2\sigma_\mu^2} \end{aligned}$$

On this surface this looks a bit complicated as a function of μ, but there is a trick: Notice that all the dependence on μ comes from the exponent. We can rewrite this as

$$A\mu^2 + B\mu + C$$

where

$$\begin{aligned} A &= \frac{1}{2}\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_\mu^2}\right) \\ B &= -\frac{\sum_i Y_i}{\sigma^2} - \frac{\mu_0}{\sigma_\mu^2} \\ C &= \frac{\sum_i Y_i^2}{2\sigma^2} + \frac{\mu_0^2}{\sigma_\mu^2} \end{aligned}$$

If we factor this quadratic equation, we find that is can be written

$$A(\mu - B/2A)^2 + \text{const.}$$

We don't care what the constant terms is, since the it is the first term which tells us the mean and standard deviation of μ. Now observe that

$$-\frac{B}{2A} = \bar{Y}\frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma^2/N} + \mu_0\frac{\sigma^2/N}{\sigma_\mu^2 + \sigma^2/N} \equiv \mu_b$$

and

$$\frac{1}{A} = 2\frac{\sigma_\mu^2\sigma^2/N}{\sigma_\mu^2 + \sigma^2/N} \equiv 2\sigma_b^2$$

In particular, we can deduce that

$$\mu|D \sim \text{Normal}\,(\mu_b,\sigma_b)$$

where μ_b and σ_b^2 defined above are the mean and variance of the posterior distribution.

- Importantly, for the Normal distribution the mean and mode are the same, so there is really only one choice for the point estimate of a parameter which has a Normal posterior.

1.2 Bayesian inference for linear regression models

- We now discuss Bayesian inference for the general linear regression model with features. Let

$$f(x) = \sum_{i=1}^{K} \beta_i \phi_i(x)$$

and suppose that our priors are

$$\beta_i \sim \text{Normal}(0, \tau_i^2).$$

We will assume that $E[\phi_i(X)] = 0$. We will also assume, for simplicity, that σ **is known. Hence, we do not need to consider estimating it.** This simplifies the conceptual picture considerably.

- In this case, the Posterior distribution of the β_i can be understood analytically. It turns out that the marginal posterior distribution of each β_i is again Normal – thus Normal priors on β_i are conjugate priors!. The marginal mean of each β_i are determined by the system of questions given in the following Theorem.

Theorem 1 (Posterior mean for regression coefficients)**.** *Define a $K \times K$ matrix Ω called the empirical covariance matrix which has entries*

$$\Omega_{i,j} = N\overline{\phi_i(X)\phi_j(X)} = \sum_{k=1}^{N}\sum_{z=1}^{N} \phi_i(X_k)\phi_j(X_z)$$

and let $\bar{\beta}_i$ denote the posterior mean of β_i; that is,

$$\bar{\beta}_i = E[\beta_i | D].$$

Then $\bar{\beta}_1, \ldots, \bar{\beta}_K$ satisfy the K equations

$$\left(\frac{\sigma}{\tau_i}\right)^2 \bar{\beta}_i + \sum_{j=1}^{K} \Omega_{i,j}\bar{\beta}_j = \sum_{j=1}^{N} \phi_i(X_j)Y_j \tag{4}$$

We make a few remarks on this Theorem:

- If $N > K$, then $\bar{\beta}_i$ actually have a solution. In-fact, depending on Ω, there may not be a solution. For our discussion, we will assume there is however.
- Under the assumption that $E[\phi_i(X)] = 0$, the entries of $\Omega_{i,j}$ approximate the covariances between the predictor features. Moreover, if we take the predictors to be drawn from some distribution and average over the data (both X and Y), we get

$$\frac{E[\Omega_{i,j}]}{N} = \text{cov}(\phi_j(X), \phi_i(X)).$$

and

$$\frac{1}{N}\sum_{j=1}^{N} E[\phi_i(X_j)Y_j] = \text{cov}(Y, \phi_i(X)).$$

Hence, we have the "math world" version of Equations 4, which is obtained by averaging over the data:

$$\frac{1}{N}\left(\frac{\sigma}{\tau_i}\right)^2 E[\bar{\beta}_i] + \sum_{j=1}^{K} \text{cov}(\phi_i(X), \phi_j(X))E[\bar{\beta}_j] = \text{cov}(Y, \phi_i(X)).$$

In the limit $N \to \infty$ and with $K = 2$, we obtain a system of equations recognizable from Week 5 notes for the regression coefficients in the two predictor model in terms of the covariances. In this context, the role of the regression coefficients (which we previously took to be fixed numbers), is now played by the averages $E[\bar{\beta}_j]$, which are the average values of β_j with respect to both the posterior and the data.

- Notice that the first term in Equation 4 vanishes as either $\tau_i \to \infty$ (very broad priors) or $\sigma \to 0$ (no variance in Y conditional on the predictors). In both cases, the vales of $\bar{\beta}_i$ are entirely determined by the data, and the priors play no role. When $\tau_i \to 0$ or $\sigma \to \infty$, the priors entirely determine determine $\bar{\beta}_i$ and in fact $\bar{\beta}_i = 0$.
- The first term in Equation 4 is an example regularization.

- Theorem 1 can be elegantly stated using matrices (see Linear algebra review note for intro to matrix multiplication). This is import if we want to implement these computations in Python. Observe that one way to construct $\Omega_{i,j}$ is to define the matrices

$$A = \begin{bmatrix} | & | & & | \\ \phi_1(X) & \phi_2(X) & \cdots & \phi_K(X) \\ | & | & & | \end{bmatrix}, \quad \Lambda_0 = \begin{bmatrix} \tau_1 & 0 & \dots & 0 \\ 0 & \tau_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \tau_K \end{bmatrix}$$

Then $\Omega = A^T A$. Now define another matrix $K \times K$ matrix

$$M = \Omega + \sigma^2 \Lambda_0^{-1}$$

Then Equations 4 can be written as

$$M\bar{\beta} = A^T y$$

where $\bar{\beta}$ and y are respectively a K-vector and N-vector

$$\bar{\beta} = \begin{bmatrix} \bar{\beta}_1 \\ \vdots \\ \bar{\beta}_K \end{bmatrix}, \quad y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix}.$$

Then $\bar{\beta}$ satisfies

$$M\bar{\beta} = A^T y \implies \bar{\beta} = \left(A^T A + \sigma^2 \Lambda_0^{-1}\right)^{-1} A^T y$$

Example 4 (Calculating posterior mean in python)**.** Consider the model with two predictors:

$$f(x) = \beta_1 X_1 + \beta_2 X_2$$

Question: Generate some simulated data and compute the posterior means of β for different choices of τ.

Solution: See colab notebook.

1.2.1 Posterior prediction

- We will mostly focus on how we can use the posterior estimates to make predictions, although there is an entirely different set of questions we could ask concerning model assessment and hypothesis testing.Recall that $\hat{y}(x, D)$ is defined as our prediction, using the regression model fitted to the data D, of $E[Y|X = x]$. In the Bayesian context, we can define the posterior predictive variable $Y|X, D$ as the random representing

the distribution of our response variable conditioned on the data. We can obtain this distribution by replacing the β_is in

$$Y = \sum_{i=1}^{K} \beta_i \phi_i(X)$$

with samples from the posterior distribution. In the Bayesian framework we set

$$\hat{y}(x, D) = E[Y|D, X] = \sum_{i=1}^{K} E[\beta_i|D]\phi_i(X) = \sum_{i=1}^{K} \bar{\beta}_i \phi_i(X)$$

We can write this in matrix notation by defining the vector

$$a = \begin{bmatrix} \phi_1(X) \\ \vdots \\ \phi_K(X) \end{bmatrix}.$$

Then

$$\hat{y}(x, D) = a^T \bar{\beta} = a^T M^{-1} A^T y \tag{5}$$

Example 5 (Posterior predictions for a regression model)**.** Take the Fourier model

$$f(x) = \beta_0 + \sum_{i=1}^{K} \beta_i \sin\left(\frac{2\pi i x}{L}\right) + \alpha_i \cos\left(\frac{2\pi i x}{L}\right).$$

Implement a function to compute posterior predictions with different priors τ on the coefficients. In this example, we will fit this model to data generated by adding noise to the function

$$f_{\text{true}}(x) = \sin(2\pi 3x) + 5\sin(2\pi 6x^2) + 20x^2$$

with $x \in [0, 1]$. Note that for $K < \infty$, we can not pick any values of the β_is and α_is such that $f = f_{\text{true}}$.

We will assume that β_i and α_i have the same prior distribution:

$$\beta_i \sim \text{Normal}(0, \tau_i)$$
$$\alpha_i \sim \text{Normal}(0, \tau_i)$$

Question: Compute the posterior mean $\hat{y}(x, D)$ for the following choices of priors and $K = 50$.

(a) All τ_i are the same, $\tau_i = \tau_0$.

(b) The τ_i decay with i as $\tau_i = \tau_0/i$

In each case, experiment with different choices of τ_0 and pay attention to whether the model is overfitting (high variance, low bias) or under fitting (low variance, high bias).

Solution: See colab notebook.

1.3 Regularization view of priors

- Recall that in standard least squares linear regression, the regression coefficients come from minimizing the sum of squared residuals. That is, they are the values that minimize the function

$$L(\beta) = \sum_{j=1}^{N} (\hat{y}(X_j, D) - Y_j)^2$$

This is an example of a loss function, which is simply something we want to minimize to solve our inference problem. This means that the derivative of L with respect to each β_i is zero.

$$\begin{aligned}
\frac{\partial}{\partial \beta_i} L(\beta) &= \sum_{z=1}^{N} 2(\hat{y}(X_z, D) - Y_z) \frac{\partial}{\partial \beta_i} \hat{y}(X_j, D) \\
&= \sum_{z=1}^{N} 2(\hat{y}(X_z, D) - Y_z) \phi_i(X_z) \\
&= 2 \sum_{z=1}^{N} \sum_{j=1}^{K} \beta_j \phi_i(X_z) \phi_j(X_z) - 2 \sum_{z=1}^{N} Y_z \phi_i(X_z) \\
&= 2 \sum_{z=1}^{K} \beta_j \sum_{z=1}^{N} \phi_i(X_z) \phi_j(X_z) - 2 \sum_{z=1}^{N} Y_z \phi_i(X_z) \\
&= 2 \sum_{j=1}^{K} \beta_j \Omega_{i,j} - 2 \sum_{z=1}^{N} Y_z \phi_i(X_z)
\end{aligned}$$

Setting

$$\frac{\partial}{\partial \beta_i} L(\beta) = 0$$

leads to Equation 4 without the prior term.

- The regularization view of Equation 4 is that, instead of thinking of the additional term as coming from priors, we think about adding a term to our loss function which penalizes models with too much flexibility. A common choice is

$$L(\beta) = \sum_{j=1}^{N} (\hat{y}(X_j, D) - Y_j)^2 + \lambda \sum_{j=1}^{K} \beta_j^2$$

This is called ridge regression and λ is a parameter controlling how large a penalty we place on large values of β_j. If you compute the partial derivatives and equate it to zero, you will find that the equations β_j satisfy are exactly Equation 4 with $\sigma/\tau_i = \sqrt{\lambda}$. More generally, we could use the loss function

$$L(\beta) = \sum_{j=1}^{N} (\hat{y}(X_j, D) - Y_j)^2 + \sum_{j=1}^{K} \left(\frac{\sigma}{\tau_i} \right)^2 \beta_j^2$$

we would obtain exactly Equation 4.

1.4 The kernel trick

- The problem of computing $\hat{y}(x, D)$ is an example of smoothing, or interpolating. Smoothing refers to the situation where we are given noisy measurements of a function $f(x)$ and want to "smooth out the noise". One way to do this is take the weighted averaging of neighboring values of y. This following examples illustrates how regression modeling is related to smoothing.

Example 6 (Orthogonal empirical covariance matrix)**.** Consider the special case where the empirical covariance matrix is diagonal – that is, $\Omega_{i,j} = 1$ if $i = j$ and 0 if $i \neq j$. Recall that this is the case when we use the fourier model and the X_i are equally spaced.
Question:

(a) Write down a formula for $\hat{y}(x, D)$ in this case. In particular, show how to express $\hat{y}(x, D)$ in the form

$$\hat{y}(x, D) = \sum_{j=1}^{N} Y_j \kappa(x, X_j) \tag{6}$$

for some function $\kappa(x, x')$.

(b) What does this function $\kappa(x, x')$ look like when $f(x)$ is a Fourier model?

Solution:

(a) In this case,

$$\bar{\beta}_i = \frac{1}{1 + \left(\frac{\sigma}{\tau_i}\right)^2} \sum_{j=1}^{N} \phi_i(X_j) Y_j$$

and hence

$$\begin{aligned}\hat{y}(x, D) &= \sum_{i=1}^{K} \frac{\phi_i(x)}{1 + \left(\frac{\sigma}{\tau_i}\right)^2} \sum_{j=1}^{N} \phi_i(X_j) Y_j \\ &= \sum_{j=1}^{N} Y_j \left(\sum_{i=1}^{K} \frac{\phi_i(x)}{1 + \left(\frac{\sigma}{\tau_i}\right)^2} \phi_i(X_j) \right)\end{aligned}$$

so

$$\kappa(x, x') = \sum_{i=1}^{K} \frac{\phi_i(x)\phi_i(x')}{1 + \left(\frac{\sigma}{\tau_i}\right)^2}$$

(b) For the Fourier model with $L = 1$,

$$\kappa(x, x') = \sum_{i=1}^{K} \frac{\sin(2\pi i x)\sin(2\pi i x') + \cos(2\pi i x)\cos(2\pi i x')}{1 + \left(\frac{\sigma}{\tau_i}\right)^2}$$

Using the identity

$$\cos(x)\cos(y) + \sin(x)\sin(y) = \cos(x - y)$$

and the fact that $\cos(x - y) = \cos(y - x)$, we get can rewrite this as

$$\kappa(x, x') = \eta(|x - x'|) = \sum_{i=1}^{K} \frac{\cos(2\pi i |x - x'|)}{1 + \left(\frac{\sigma}{\tau_i}\right)^2}$$

- We actually don't need the empirical covariance matrix to be diagonal to "kernalize" our predictions

(meaning express $\hat{y}(x, D)$ in the form of Equation 6. Equation 5 also has this form, although we can not determine the kernel without inverting a very large matrix.