6.7900 Fall 2024: Lecture Notes 6

1 Recap

Last time we talked about potential problems with MLE and empirical risk minimization, especially for linear regression.

A possible issue is that you pick out one exact estimate, but you might want to express uncertainty. For example, sometimes there are many solutions to a linear regression problem, and you would like to express some uncertainty about your estimate.

Mathematically, this could arise when you have an ill-conditioned matrix due to collinearity (fewer examples than parameters).

2 The Model

We analyze the following model:

$$y^{(i)} \sim \mathcal{N}(\theta^T x^{(i)}, \sigma^2),$$
$$\sigma^2 \text{ known.}$$

Here, we have

- $(x^{(i)}, y^{(i)})$ are the data points.
- θ is the parameter we want to estimate.
- σ^2 is the noise variance.

The assumption of a known σ^2 is not that large of a deal. If it were not known, we could use Bayesian techniques to model our uncertainty about it, and update our belief over time based on the data we get, just as we are doing with the θ parameters. It's not any conceptually harder, but gets kind of complicated to write down. To keep the example somewhat simple, we'll stay with assuming it's known.

From this model, we have:

- Prior (our initial belief of θ):

$$p(\theta) = N(\theta|\mu_0, \Sigma_0).$$

- Posterior (our belief of θ after observing data):

$$\begin{aligned} p(\theta|\{y^{(i)}\}) &= N(\theta|\mu_N, \Sigma_N) \\ \Sigma_N &= (\Sigma_0^{-1} + \sigma^{-2} X^T X)^{-1}, \\ \mu_N &= \Sigma_N(\Sigma_0^{-1}\mu_0 + \sigma^{-2} X^T y). \end{aligned}$$

- Posterior predictive (our belief of the next data point after observing data):

$$\begin{aligned} p(y^{(N+1)}|y^{(1)}, \ldots, y^{(N)}) &= \mathcal{N}(y^{(N+1)}|\mu_{N+1}, V_{N+1}^2), \\ \mu_{N+1} &= \mu_N^T x^{(N+1)}, \\ V_{N+1} &= \sigma^2 + \left(x^{(N+1)}\right)^T \Sigma_N x^{(N+1)}. \end{aligned}$$

Exercise: For the posterior mean and covariance matrix, we invert a matrix. Show that we that invertibility is guaranteed.

An important note is we have a measure of uncertainty for θ, in the sense that we estimate the variance of the posterior.

In the following few sections, we will do some demos to gain intuition about the posterior, the posterior predictive, and the effect of the prior. There are visualizations here that will not be in the notes, so it is recommended to watch the lecture.

3 Demo A

3.1 Initial Setup

We generate data according to the following model:

$$\begin{aligned} x^{(n)} &\sim U(-1, 1), \\ y^{(n)} &= \mathcal{N}(\theta_{1,true} + \theta_{2,true} x^{(n)}, \sigma^2), \\ \theta_{1,true} &= -0.3, \theta_{2,true} = 0.5, \sigma = 0.2. \end{aligned}$$

Our goal is to estimate θ. We have the following assumptions:

- Our prior is Gaussian:

$$p(\theta) = \mathcal{N}(\theta|0_D, \sigma_0^2 I_{D\times D}), \sigma_0^2 = 1$$

- The underlying model is linear with known noise:

$$y^{(n)} \sim \mathcal{N}(\theta_1 + \theta_2 x^{(n)}, \sigma^2).$$

We conduct the following analysis:

- Zero observations: the posterior is the prior.
- One observation: the posterior represents that lines should roughly go through the observation.
- More observations: the posterior becomes more concentrated. The sampled lines from the posterior get more concentrated too.

3.2 Posterior Predictive

Now we analyze the predictive posterior too. We fix the x value we wish to predict at. We conduct the following analysis:

- One observation: the posterior predictive has large uncertainty.
- More observations: the posterior predictive still has large uncertainty, but slightly less.

Crucially, we note that our posterior uncertainty goes down with more observations, but our predictive posterior uncertainty decreases but tapers to a constant. This is interesting, because our belief about θ is very certain, but our belief about a new data point is not.

We conduct a sense check on $D = 1$. We have a posterior variance of

$$\sigma_N^2 = \left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2} \sum_{n=1}^{N} \left(x^{(n)} \right)^2 \right)^{-1}.$$

We have a posterior predictive variance of

$$V_n = \sigma^2 + \left(x^{(N+1)} \right)^2 \sigma_N^2.$$

The posterior variance always goes down with more observations and decreases to 0. However, the posterior predictive variance is lower bounded by σ^2. This makes sense though; the data is drawn with an inherent noise of σ^2. No matter what, we always have this unavoidable noise, which is represented in the posterior predictive.

4 Demo B

We keep the same model as Demo A, but instead of sampling $x^{(n)} \sim U[-1, -1]$ we sample $x^{(n)} = 0.2$. In other words, we only observe data points with a fixed x value. We conduct the following analysis:

- One observation: the posterior is the same as Demo A.
- More observations: the posterior informs us only that the line will go through the center of the observed points. The observed points consist of a vertical line at $x = 0.2$.

The posterior represents that we are certain that the line will go through the center of the cloud of points observed at $x = 0.2$. However, it doesn't know anything else about the line. Further, note that the posterior is still concentrated towards $(0, 0)$ because we have the prior that θ is normal centered at $(0, 0)$.

This uncertainty also represents the collinearity in our dataset, which is translated into a large uncertainty in the posterior. In the MLE setting, this would be represented by inverting an ill-conditioned matrix.

5 Demo C

We assume our assumption of a linear model is wrong. We generate data according to the following model:

$$x^{(n)} \sim U(-1, 1),$$
$$y^{(n)} = \mathcal{N}(0.5 - (x^{(n)})^2, \sigma^2).$$

We conduct the following analysis:

- Many observations: the posterior variance will go down again, because more data always reduces uncertainty. The posterior predictive also looks the same as before.

It is interesting that the posterior variance goes down, but we clearly see that our model is a bad fit for the data.

6 Recap of Lessons

Instead of returning an error when there is collinearity like MLE, the Bayesian approach gives us a distribution over possible θ. This is a more informative way to represent uncertainty.

It is common to see publications discuss different uncertainties.

- Aleatoric uncertainty: intrinsic randomness.
- Epistemic uncertainty: uncertainty due to lack of knowledge.

Consider the example of shuffling a deck of cards, and we ask if the top card is an ace. Before we shuffle, uncertainty is aleatoric. After shuffling but before looking at the top card, the uncertainty is epistemic. This is because before, there is some intrinsic randomness, but after shuffling, the uncertainty is just due to lack of knowledge.

There are philosophical nuances here: maybe if you had a perfect physics model of the physics of shuffling, you'd say there was no intrinsic randomness in the process, and all the uncertainty was epistemic, though, so it's hard to make this distinction completely crisp.

One other way these terms relate to what we've been doing is that, in the posterior predictive distribution, we can see σ^2 as representing the aleatoric uncertainty and (that second term which I'm not pasting in here) as representing our epistemic uncertainty.

7 MAP and Regularization

Now, what if you just want a single regression line and not a whole distribution? An option is the posterior mean or the MAP. For the linear regression model, this is

$$\begin{aligned}\hat{\theta}_{\text{MAP}} &= \text{argmax}_\theta p(\theta|\mathcal{D}) = \text{argmin}_\theta - \log p(\theta|\mathcal{D}) \\ &= \text{argmin} + \theta \left\{ (X\theta - Y)^T (X\theta - Y) + \frac{\sigma^2}{\sigma_0^2} \theta^T \theta \right\}.\end{aligned}$$

But this is just the MLE solution, except we added an l_2/ridge penalty/regularizer of $\frac{\sigma^2}{\sigma_0^2}\theta^T\theta$. An intuition behind this penalty term is that we penalize large θ, which follows our intuition that θ should be centered at 0 from our prior. This solution is therefore called ridge regression; this is a regularized verison of the MLE. The term ridge penalty has historical origins. The closed form solution is

$$\begin{aligned}\lambda &= \frac{\sigma^2}{\sigma_0^2}, \\ \hat{\theta}_{\text{MAP}} &= (X^T X + \lambda I_{D\times D})^{-1} X^T Y.\end{aligned}$$

λ intuitively controls the strength of the ridge penalty. When $\lambda \to 0$, this becomes the OLS solution. When $\lambda \to \infty$, this becomes the prior mean, which is 0. In a sense, we avoid the potential issues with inverting ill-condition matrices when performing MLE by choosing a particular option among options "equally good" in MLE. Note that this solution may offer better generalization capabilities than vanilla MLE.

8 A Note on Features

Linear models can be very flexible given non-trivial features. We've been considering $h(x) = \theta_1 x_1 + \cdots + \theta_D x_D = \theta^T x$. But we could've taken $h(x) = \theta_1 \phi_1(x) + \cdots + \theta_D \phi_D(x) = \theta^T \phi(x)$. Now x can be any dimension, and D is the dimension of the features $\phi(x)$. We could create features $x \in \mathbb{R}, \phi(x) = [1, x, x^2]^T$, which betters fit our Demo C. More generally, for $x \in \mathbb{R}^{D_x}, \phi(x)$ could collect polynomials of degree r or smaller, which is polynomial regression.

For any fixed $\phi(x)$, all the math we did can be done as before, just replacing $\phi(x)$ in place of x.