Course 67564 2022/23

Bayesian Machine Learning

Recitation 5: Evidence Function and Predictive Distribution

Prof. Yair Weiss *TA: Roy Friedman*

1 Evidence Function

As we have previously discussed, there are many possible basis functions $h(\cdot)$ we can use to fit the linear regression model, and it is not always so simple to determine which set of basis functions is the correct one to use. On one hand, if we use a very expressive set of basis functions, or a very large one, then the model will easily fit the *training* data, but will probably give very inaccurate predictions for unseen data points. On the other hand, if we use a model that is too simplistic, then we will end up missing all of the data points.

This is exactly the dilemma represented in figure 1; on the left, it is fairly obvious that the straight line should be chosen, although the 9th order polynomial fits the data better. On the other hand, the graph on the right shows exactly the opposite - the 9th order polynomial intuitively looks like it explains the data better than the linear function. However, in both cases the 9th order polynomial has much higher likelihood. So how can we choose which of the basis functions is a better fit for the data?

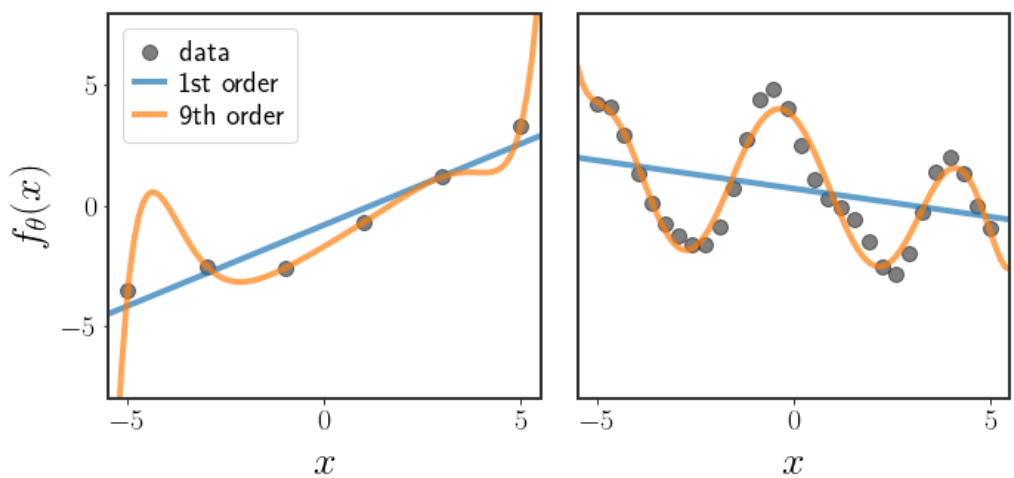

Figure 1: an example of the dilemma of choosing which model should be used. In both plots, a linear function and a 9th order polynomial are fitted to the data. In each case, how can we choose which of the basis functions should be used for linear regression?

The *evidence function*[1] (also called the *marginal likelihood*, since we marginalize the parameters out of the distribution) is a way for us to intelligently choose which parameterization to use. The idea behind the evidence function is to "integrate out" the specific values of the parameters θ and to see how probable the data set is under our parameterization. Suppose we have a prior $p(\theta|\Psi)$ that is dependent on some parameters Ψ. Then:

$$\begin{aligned} p(\mathcal{D}|\Psi) &= \int p(\mathcal{D},\theta|\Psi)\, d\theta \\ &= \int \underbrace{p(\mathcal{D}|\theta)}_{\text{likelihood}} \cdot \underbrace{p(\theta|\Psi)}_{\text{prior}}\, d\theta \end{aligned} \tag{1.1}$$

[1] Bishop section 3.4

The way it is written at the moment may be a bit confusing. Up until now (and from now on, as well), we wrote the prior as $p(\theta)$, but suddenly we're adding the conditioning on Ψ - why? When we define a prior, we usually have to choose a distribution for the prior. Often, this distribution has parameters that define it; for instance, if the prior is a Gaussian, then the parameters are the specific μ_0 and Σ_0 we chose. In our new notation $\Psi = \{\mu_0, \Sigma_0\}$, and we want to compare between different possible Ψs.

Simple Bayesian linear regression

As an example, suppose we assume that:

$$y = \theta x \tag{1.2}$$

Furthermore, we assume that we have two priors on θ given by:

$$\theta \sim \mathcal{N}(\mu_1, \Sigma_1) \tag{1.3}$$

$$\tilde{\theta} \sim \mathcal{N}(\mu_2, \Sigma_2) \tag{1.4}$$

In other words, we have two competing priors, from which we want to choose only one. We can then calculate:

$$p(y|\mu_i, \Sigma_i) = \int p(y|\theta)\, p(\theta|\mu_i, \Sigma_i)\, d\theta \tag{1.5}$$

for $i \in \{1, 2\}$. If we calculate this probability for both μ_1 and μ_2, then we will get a value that tells us how likely the training data y is under each of these different assumptions. That is, instead of asking how probable y is under a specific value of θ (which is just the likelihood), this is like asking how probable y is when averaging out the values of θ, *given the parameterization* of μ_i and Σ_i.
Of course, the priors could assume different basis functions, as in:

$$f_1(x) = \theta x \qquad \theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta) \tag{1.6}$$

$$f_2(x) = \beta_0 + \beta_1 x \qquad \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \sim \mathcal{N}(\mu_\beta, \Sigma_\beta) \tag{1.7}$$

and we want to choose between $\Psi_\theta = \{\mu_\theta, \Sigma_\theta\}$ and $\Psi_\beta = \{\mu_\beta, \Sigma_\beta\}$. Notice that the basis functions we assume are themselves *implicitly* compared when we do this, simply since the basis functions are part of the assumptions we made when we chose our prior.

1.1 Finding the Evidence

Suppose that our prior is described, as above, by:

$$\theta \sim p(\theta|\ \Psi) \tag{1.8}$$

where Ψ are some parameters. We want to find the value of $p(\mathcal{D}|\ \Psi)$ - the evidence for seeing the data under this parameterization Ψ. Recall from Bayes' law:

$$p(\theta|\ \mathcal{D}, \Psi) = \frac{p(\theta, \mathcal{D}|\Psi)}{p(\mathcal{D}|\Psi)} \tag{1.9}$$

$$\Leftrightarrow p(\mathcal{D}|\Psi) = \frac{p(\theta, \mathcal{D}|\Psi)}{p(\theta|\ \mathcal{D}, \Psi)} \tag{1.10}$$

$$\Leftrightarrow p(\mathcal{D}|\Psi) = \frac{p(\mathcal{D}|\theta)\, p(\theta|\Psi)}{p(\theta|\ \mathcal{D}, \Psi)} \tag{1.11}$$

This is true for *every* choice of θ! A common way to find the evidence function is by plugging $\hat{\theta}_{\text{MAP}}$ into the expression, which gives:

$$\Leftrightarrow p(\mathcal{D}|\Psi) = \frac{p\left(\mathcal{D}|\hat{\theta}_{\text{MAP}}\right) p\left(\hat{\theta}_{\text{MAP}}|\Psi\right)}{p\left(\hat{\theta}_{\text{MAP}}|\ \mathcal{D}, \alpha\right)} \tag{1.12}$$

Usually, the numerator is either known or pretty simple to calculate, while the denominator is quite hard to find. In such cases, the denominator is approximated in some manner in other to find the evidence. Luckily for us, the denominator is easy to calculate in the case of Bayesian linear regression with a Gaussian prior.

1.2 Evidence for Bayesian Linear Regression

In standard Bayesian linear regression, the posterior is a Gaussian. We can utilize this knowledge to find a more specific formula for the evidence. Notice that:

$$p\left(\hat{\theta}_{\text{MAP}} \mid y, \Psi\right) = \max_{\theta} p\left(\theta \mid \mathcal{D}, \Psi\right) \tag{1.13}$$

$$= \max_{\theta} \frac{1}{\sqrt{(2\pi)^d \left|\Sigma_{\theta|\mathcal{D}}\right|}} e^{-\frac{1}{2}\left(\theta-\mu_{\theta|\mathcal{D}}\right)^T \Sigma_{\theta|\mathcal{D}}^{-1}\left(\theta-\mu_{\theta|\mathcal{D}}\right)} \tag{1.14}$$

This maximum is attained at $\theta = \mu_{\theta|\mathcal{D}}$, where the whole term in the exponent is equal to 1, so we're left with:

$$p\left(\hat{\theta}_{\text{MAP}} \mid y, \Psi\right) = \frac{1}{\sqrt{(2\pi)^d \left|\Sigma_{\theta|\mathcal{D}}\right|}} \tag{1.15}$$

Plugging this into equation 1.12, we have:

$$p\left(y|\Psi\right) = (2\pi)^{N/2} \left|\Sigma_{\theta|\mathcal{D}}\right|^{1/2} p\left(y|\hat{\theta}_{\text{MAP}}\right) p\left(\hat{\theta}_{\text{MAP}}|\Psi\right) \tag{1.16}$$

But we can be even more specific. Remember that for a Gaussian prior $\hat{\theta}_{\text{MAP}} = \mu_{\theta|D}$ and $\Psi = \{\mu, \Sigma\}$, so what we actually found is:

$$p\left(y|\mu, \Sigma\right) = (2\pi)^{N/2} \left|\Sigma_{\theta|\mathcal{D}}\right|^{1/2} \mathcal{N}\left(\mu_{\theta|\mathcal{D}} \mid \mu, \Sigma\right) \mathcal{N}\left(y \mid H\mu_{\theta|\mathcal{D}}, I\sigma^2\right) \tag{1.17}$$

1.3 Equivalent Derivation

The above derivation allows us to calculate the actual value of the evidence quickly, but it may be a bit harder to understand what is going on in this form. An equivalent way to find the evidence is to find $p\left(\mathcal{D}|\Psi\right)$ directly, from the definition. Recall that we modeled linear regression according to:

$$y = H\theta + \eta \quad \eta \sim \mathcal{N}\left(0, I\sigma^2\right) \tag{1.18}$$

If we marginalize θ out of the above equation, it will still be an exponent of something that is quadratic in y, so it will be a Gaussian. So we just need to find the mean and covariance in order to find the exact form of the Gaussian:

$$\begin{aligned} \mathbb{E}\left[y\right] &= \mathbb{E}\left[H\theta + \eta\right] \\ &= H\mathbb{E}\left[\theta\right] + \mathbb{E}\left[\eta\right] \\ &= H\mu \end{aligned} \tag{1.19}$$

The expectation is always the easiest part, but in this case the covariance isn't much harder to find:

$$\begin{aligned} \mathbb{E}\left[yy^T\right] &= \mathbb{E}\left[\left(H\theta + \eta\right)\left(H\theta + \eta\right)^T\right] \\ &= \mathbb{E}\left[H\theta\theta^T H^T + \eta\theta^T H^T + H\theta\eta^T + \eta\eta^T\right] \\ &= \mathbb{E}\left[H\theta\theta^T H^T\right] + \mathbb{E}\left[\eta\eta^T\right] \\ &= H\mathbb{E}\left[\theta\theta^T\right] H^T + I\sigma^2 \\ &= H\left(\Sigma + \mu\mu^T\right) H^T + I\sigma^2 \end{aligned} \tag{1.20}$$

$$\Rightarrow \text{cov}\left[y\right] = \mathbb{E}\left[yy^T\right] - \mathbb{E}\left[y\right]\mathbb{E}\left[y^T\right] = H\Sigma H^T + I\sigma^2 \tag{1.21}$$

So, the evidence function for Bayesian linear regression is actually the density of the following Gaussian at the point y:

$$p\left(y|\mu, \Sigma\right) = \mathcal{N}\left(y \mid H\mu,\ H\Sigma H^T + I\sigma^2\right) \tag{1.22}$$

Learning the sample noise

Notice that all this time we assumed that we know the variance of the sample noise, σ^2. This really helps simplify many of the derivations we made, but is kind of a weird assumption to make.
We can try to use a fully Bayesian approach, where we choose a prior for σ^2 and then calculate the posterior. If we try to choose a Gaussian as a prior, we quickly run into a problem - σ^2 can't be negative, but every Gaussian will have a positive density at negative values! In addition, the Gaussian distribution is symmetric, but the distribution we want to describe σ^2 with is probably very asymmetrical, with low density at values close to zero, high density later on, and a long tail for higher values. So, clearly we can't use a Gaussian as the prior for σ^2. There are distributions that match the above description, but we haven't discussed them (and won't). Also, finding their posterior is usually a bit harder than finding the posterior of a Gaussian distribution[a]. So, going fully Bayesian is needlessly complicated, at least in this case.
Instead, we can use the evidence function in order to choose the most fitting sample noise. In the notation above, we only wrote y as a function of μ and Σ, but it is obviously affected by σ^2 through the covariance:

$$\operatorname{cov}[y] = H\Sigma H^T + I\sigma^2 \tag{1.23}$$

We can define the evidence as a function of the variance as well and then choose from a closed set of values chosen ahead of time $S = \left\{\sigma_i^2\right\}_{i=1}^q$, in which case we would say:

$$\hat{\sigma}^2 = \arg\max_{\sigma^2 \in S} \mathcal{N}\left(y \,|\, H\mu,\ H\Sigma H^T + I\sigma^2\right) \tag{1.24}$$

Another option is to use gradient ascent (or another optimization algorithm) in order to find the maximum iteratively:

$$\hat{\sigma}^2_{(t)} = \hat{\sigma}^2_{(t-1)} + \epsilon \nabla_\sigma \log p\left(y \,|\, \mu, \Sigma, \sigma\right) \tag{1.25}$$

where ϵ is some learning rate. However, note that there is no guarantee that the evidence is a concave function!

[a] If you are curious, you can look at Bishop section 2.3.6 for a full derivation of the posterior under a proper prior

2 Predictive Distribution

A lot of times we don't actually care what values θ takes, we only care about the predicted values for a new data point x_n. In this case, we can just try to integrate out θ and predict y_n directly, given our prior:

$$p\left(y_n|\Psi,\mathcal{D}\right) = \int p\left(y_n|\theta\right) p\left(\theta|\Psi,\mathcal{D}\right) d\theta = \int p\left(y_n,\theta|\Psi,y\right) d\theta \tag{2.1}$$

This is called the *predictive distribution* as what we are trying to do is *predict* new values.

In the above equation we did something that would usually not be true:

$$p\left(y_n|\theta\right) p\left(\theta|\Psi,\mathcal{D}\right) \stackrel{?}{=} p\left(y_n,\theta|\Psi,\mathcal{D}\right) \tag{2.2}$$

when what we should have written was:

$$p\left(y_n,\theta|\Psi,\mathcal{D}\right) = p\left(y_n|\theta,\Psi,\mathcal{D}\right) p\left(\theta|\Psi,\mathcal{D}\right) \tag{2.3}$$

However, note that one the first assumptions about the likelihood that we made was that given θ, y_n does not depend on $\mathcal{D}$ or Ψ. This is most apparent when you write out the full distribution for linear regression:

$$y_n|\theta,\mathcal{D},\mu,\Sigma \sim \mathcal{N}\left(h\left(x_n\right)^T\theta,\ \sigma^2\right) \tag{2.4}$$

Notice how $\mathcal{D}$, μ and Σ don't appear in this term? Because of that, we are allowed to drop the dependence on the three variables, **as long as y_n is also conditioned on θ -** in any other case it would not be true.

There are multiple ways we can go about finding the predictive distribution, but we'll use a similar method to what we did in section 1.3 (the predictive distribution is even a kind of extension of the evidence). First, notice that the following equation completely describes the marginal distribution:

$$y_n = h\left(x_n\right)^T\theta|\mathcal{D} + \eta \quad \eta \sim \mathcal{N}\left(0,\sigma^2\right) \tag{2.5}$$

Now, since any linear transformation of a Gaussian is also a Gaussian, we know that y_n will be a Gaussian distribution, so:

$$\mathbb{E}\left[y_n\right] = \mathbb{E}\left[h\left(x_n\right)^T \theta|\mathcal{D} + \eta\right] = h\left(x_n\right)^T \mathbb{E}\left[\theta|\mathcal{D}\right] + \mathbb{E}\left[\eta\right] = h\left(x_n\right)^T \mu_{\theta|D} \tag{2.6}$$

and we have found the mean! The variance will be only a little bit trickier:

$$\begin{aligned}
\operatorname{var}\left[y_n\right] &= \operatorname{var}\left[h\left(x_n\right)^T \theta|\mathcal{D} + \eta\right] && (2.7)\\
&= \operatorname{var}\left[h\left(x_n\right)^T \theta|\mathcal{D}\right] + \operatorname{var}\left[\eta\right] && (2.8)\\
&= \operatorname{var}\left[h\left(x_n\right)^T \theta|\mathcal{D}\right] + \sigma^2 && (2.9)
\end{aligned}$$

where the second step is due to the fact that η and $\theta|\mathcal{D}$ are independent. Finding the variance of $\operatorname{var}\left[h\left(x_n\right)^T \theta|\mathcal{D}\right]$ seems trickier, but recall that the variance is a special case of the covariance. So if we use the definition of the covariance:

$$\begin{aligned}
\operatorname{cov}\left[h\left(x_n\right)^T \theta|\mathcal{D}\right] &= \mathbb{E}\left[h\left(x_n\right)^T \theta|\mathcal{D}\left(h\left(x_n\right)^T \theta|\mathcal{D}\right)^T\right] - \mathbb{E}\left[h\left(x_n\right)^T \theta|\mathcal{D}\right]\mathbb{E}\left[h\left(x_n\right)^T \theta|\mathcal{D}\right]^T && (2.10)\\
&= \mathbb{E}\left[h\left(x_n\right)^T \theta|\mathcal{D}\left(\theta|\mathcal{D}\right)^T h\left(x_n\right)\right] - h\left(x_n\right)^T \mu_{\theta|D}\mu_{\theta|D}^T h\left(x_n\right)^T && (2.11)\\
&= h\left(x_n\right)^T \mathbb{E}\left[\theta|\mathcal{D}\left(\theta|\mathcal{D}\right)^T\right] h\left(x_n\right) - h\left(x_n\right)^T \mu_{\theta|D}\mu_{\theta|D}^T h\left(x_n\right)^T && (2.12)\\
&= h\left(x_n\right)^T \left(\Sigma_{\theta|D} + \mu_{\theta|D}\mu_{\theta|D}^T\right) h\left(x_n\right) - h\left(x_n\right)^T \mu_{\theta|D}\mu_{\theta|D}^T h\left(x_n\right)^T && (2.13)\\
&= h\left(x_n\right)^T \Sigma_{\theta|D} h\left(x_n\right) && (2.14)
\end{aligned}$$

... and we found the variance. The full predictive distribution is given by:

$$y_n|\mu, \Sigma, \mathcal{D} \sim \mathcal{N}\left(\mu_{\theta|D}^T h\left(x_n\right), \quad h\left(x_n\right)^T \Sigma_{\theta|D} h\left(x_n\right) + \sigma^2\right) \tag{2.15}$$

Notice that this has exactly the same form as the evidence function from equation 1.22, using the basis functions around the new points (instead of the training points) and the posterior instead of the prior. We can also do this if we want to predict multiple new points. Denoting by H_n the design matrix of the new points, we want to find the *vector* y_n. Using the same steps as above, we will get:

$$y_n|\mu, \Sigma, \mathcal{D} \sim \mathcal{N}\left(H_n \mu_{\theta|D}, \quad H_n \Sigma_{\theta|D} H_n^T + I\sigma^2\right) \tag{2.16}$$