Bayesian Linear Regression

David S. Rosenberg

Abstract

Here we develop some basics of Bayesian linear regression. Most of the calculations for this document come from the basic theory of gaussian random variables. To keep the focus on the probabilistic and statistics concepts in this document, I've outsourced the calculations to another document, on basical normal variable theory.

1 Gaussian Linear Regression – Everything but Bayes

Given an input $x \in \mathbf{R}^d$, we'd like to predict the corresponding output $y \in \mathbf{R}$. In Gaussian linear regression, we assume that y is generated by first taking a linear function of x, namely $f(x) = x^T w$, for some $w \in \mathbf{R}^d$. Barber refers to $f(x)$ as the "**clean**" output. However, we don't get to observe $f(x)$ directly. In Gaussian regression, we assume that we observe $f(x)$ plus some random Gaussian noise ε. This setting is described mathematically in the expressions below:

$$\begin{aligned} f(x) &= w^T x \\ \varepsilon &\sim \mathcal{N}(0, \sigma^2) \\ y &= f(x) + \varepsilon. \end{aligned} \tag{1.1}$$

We can think of these expressions as describing how "nature" or "the world" generates a y value given an x:

1. We give Nature x. (Or some other process generates x.)
2. Nature computes[1] $f(x) = w^T x$.
3. Nature draws a random sample ε from $\mathcal{N}(0, \sigma^2)$.

[1] Nature knows w, though we (the data scientists) generally do not.

4. Nature tells us the value of $y = f(x) + \varepsilon$.

We can think of ε as the noise in our observation. The "**learning**" or "**estimation**" problem is to figure out what w is, given a **training set** $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ generated by this process.

Using basic properties of Gaussian distributions, we can write:

$$Y|x \sim \mathcal{N}(w^T x, \sigma^2). \tag{1.2}$$

We read this as "the conditional distribution of [the random variable] Y given input x is Gaussian with mean $w^T x$ and variance σ^2. Although there is no explicit reference to the "clean" output in 1.2, you can see it is just the mean of the Gaussian distribution.

Note that the model we have described makes no mention of how x is generated. Indeed, this is intentional. This kind of model is called a **conditional model**. We only describe what Y is like, given x. The x may be the output of an unknown random process or it may be chosen by a person designing an experiment. One can think about x simply as "input".

[show distribution for a single x]

[show conditional distribution for several x's (picture gaussian going vertically?)

[show scatter plot of samples from several randomly chosen x's , x's chosen uniformly at random]

So far, we have only specified the distribution for $Y \mid x$ up to a particular **family of distributions**. What does that mean? The distribution of $Y \mid x$ depends on the parameter w, which is unknown. We only know that

$$\text{Distribution}\,(Y \mid x) \in \left\{\mathcal{N}\left(w^T x, \sigma^2\right) \mid w \in \mathbf{R}^d\right\}.$$

Our goal is to be able to predict the distribution of Y for a given x (or perhaps some characteristic of this distribution, such as its expected value or standard deviation). To end up with a single distribution for $Y \mid x$, we'll have to do more. One approach is to come up with a **point estimate** for w. This means choosing a specific $w \in \mathbf{R}^d$, typically based on our training data. Coming up with a point estimate for w is the approach taken in classical or **"frequentist"** statistics. In Section 2 we a classical frequentist approach called maximum likelihood estimation.

By contrast to the frequentist approach, in the **Bayesian approach**, we treat the unknown w as a random variable. In this approach, we never settle

on a single w, but rather we end up producing a distribution on $w \in \mathbf{R}^d$, called the **posterior distribution**. We then get the distribution for $Y \mid x$ by integrating out w.

What about σ^2? Throughout this development, we assume that σ^2 is a known quantity. However, we can also treat it as another unknown parameter, in both the frequentist approach and the Bayesian approach.

[REWRITE:]We'll first discuss what is arguably the most important frequentist approach, namely maximum likelihood estimation. Then we will introduce and develop the Bayesian approach in some detail.

For the rest of this document, we will assume that we have a **training set** $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ of input/output pairs. Although we make no assumptions about how the $x_1, \ldots, x_n$ were chosen, we assume that conditioned on the inputs $x = (x_1, \ldots, x_n)$, the responses $y_1, \ldots, y_n$ are independent.

2 Maximum Likelihood Estimation

Recall from (1.2) that our model has the form $Y \mid x \sim \mathcal{N}(w^T x, \sigma^2)$. The conditional density for a single observation $Y_i \mid x_i$ is of course

$$p_w(y_i \mid x_i) \quad = \quad \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2}\right).$$

By our conditional independence assumption, we can write the joint density for the dataset $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$as

$$p_w(\mathcal{D}) \quad = \quad \prod_{i=1}^{n} p_w(y_i \mid x_i).$$

For a fixed dataset $\mathcal{D}$, the function $w \mapsto p_w(y \mid x)$ is called the **likelihood function**. The likelihood function gives a measure of how "likely" each w is to have given rise to the data $\mathcal{D}$. In **maximum likelihood estimation**, we choose the w that has maximum likelihood for the data $\mathcal{D}$. This estimator, know as the maximum likelihood estimator, or MLE, is

maximizes the likelihood frob teh bthe method of **maximum likelihoo** known as

When viewed as a function$n$ observ

= convenience,

Using our assumption of conditional independence, the joint conditional density for .

The joint conditional density for

Suppose we observed a pair (x, y) from this model. For each possible $w \in \mathbf{R}^d$, we want a measure of how "likely" w is to have produced y for the given x.

We say that the **likelihood** that

3 Bayesian Method

In the Bayesian approach, we assign a probability distribution to all unknown parameters. The distribution should represent our "**prior belief**" about the value of w. Let's consider the case of a Gaussian prior distribution on w, namely $w \sim \mathcal{N}(0, \Sigma_p)$. Now the full model for y can be written as

$$\begin{aligned} Y &= f(x) + \varepsilon \\ f(x) &= w^T x \\ \varepsilon &\sim \mathcal{N}(0, \sigma_n^2) \\ w &\sim \mathcal{N}(0, \Sigma_p). \end{aligned}$$

Here we assume that both σ_n^2 and Σ_p are known. Note that we now have a fully specified probability distribution for $Y \mid x$ – there are no "unknown parameters". We now consider w to be an unobserved random variable. [Mathematically, it has the same status as ε.] [While previously w was an unknown parameter, now it is an unobserved random variable with a specific probability distribution.]

3.1 Posterior for a single observation [Barber 18.1.1]

[Plan – possibly simplify this treatment by giving a "completing the square identity" that we can plug into

Now suppose we observe a single input/output pair from this model: $\mathcal{D} = \{(x, y)\}$. The likelihood of any particular $w \in \mathbf{R}^d$ for the data $\mathcal{D}$ is

given by

$$\begin{aligned} p(y \mid x, w) &= \mathcal{N}(y\,;\, w^T x, \sigma_n^2) \\ &= \frac{1}{\sigma_n\sqrt{2\pi}} \exp\left(-\frac{\left(y - w^T x\right)^2}{2\sigma_n^2}\right). \end{aligned}$$

Now, the data $\mathcal{D}$ has given us some information about the relationship between y and x. This allows us to update our belief about w. Mathematically, this amounts to computing the distribution of w, conditional on the data. This distribution is called the **posterior distribution**:

$$\begin{aligned} p(w \mid \mathcal{D}) &= p(w \mid y, x) \\ &= \frac{p(y \mid w, x)p(w)}{p(y \mid x)} \text{ (using fact that } w \text{ is independent of } x) \\ &= \frac{\mathcal{N}(y\,;\, w^T x, \sigma_n^2)\mathcal{N}(w\,;\, 0, \Sigma_p)}{p(y \mid x)}. \end{aligned}$$

The denominator $p(y \mid x)$ is called the **marginal likelihood**. Note that it is independent of the weights w. [At this point would be good to analyze this expression... nothing that since the LHS is a probability distribution in w, so is the RHS. Since $p(y \mid x)$ is independent of w, it is just a proportionality constant. We can recompute it anytime we want by integrating the RHS.

$$p(w \mid \Gamma) \quad \propto \quad f(w, \Gamma),$$

where Γ can be a set of many other parameters or variables. Then to get the proportionality constant, we only have to integrate the RHS over w. So we get

$$k(\Gamma) = \int_w f(w, \Gamma),$$

and the full expression is

$$p(w \mid \Gamma) = k(\Gamma) f(w, \Gamma).$$

Whenever you want to use proportionality, rather than equality, make sure you are keeping track of what is the variable you need to integrate over to recover the proportionality constant.]

To get the proportionality constant, we simply integrate the RHS over w. Note that the proportionality constant may depend on the other parameters

[Move this to a later section] To compute it, we need to introduce the weight and integrate it out:

$$\begin{aligned} p(y \mid x) &= \int p(y, w \mid x)\, dw \\ &= \int p(y \mid w, x) p(w)\, dw \\ &= \int \mathcal{N}(y\,;\, w^T x, \sigma_n^2) \mathcal{N}(w\,;\, 0, \Sigma_p)\, dw. \end{aligned}$$

The marginal likelihood $p(y \mid x)$ has a very interesting interpretation, and we will return to this later.

Note that $p(w \mid \mathcal{D})$ has the product of two Gaussian densities $\mathcal{N}(y\,;\, w^T x, \sigma_n^2)\mathcal{N}(w\,;\, 0, \Sigma_p)$ both in the numerator, and in the integral of the marginal likelihood. It turns out that the product of these Gaussian densities can be rewritten as something proportional to a single Gaussian density, which makes it much easier to work with. Below we derive that

$$p(w \mid \mathcal{D}) = \mathcal{N}(w\,;\, \mu_\pi, \Sigma_\pi),$$

where

$$\begin{aligned} \mu_\pi &= v = M^{-1} y x = y \left(x x^T + \sigma_n^2 \Sigma^{-1}\right)^{-1} x \\ \Sigma_\pi &= \sigma_n^2 M^{-1} = \sigma_n^2 \left(x x^T + \sigma_n^2 \Sigma^{-1}\right)^{-1} = \left(\sigma_n^{-2} x x^T + \Sigma^{-1}\right)^{-1} \end{aligned}$$

$$\begin{aligned} \mathcal{N}(y\,;\, w^T x, \sigma_n^2)\mathcal{N}(w\,;\, 0, \Sigma_p) &= \alpha \mathcal{N}(w\,;\, \mu_1, \Sigma_1) \\ & \quad k' \exp\left(-\frac{1}{2}(w - v)^T \sigma_n^{-2} M (w - v)\right) \end{aligned}$$

where

$$\begin{aligned} \Sigma_1^{-1} &= \sigma_n^{-2}\left(x x^T + \sigma_n^2 \Sigma^{-1}\right) \\ \mu &= y \left(x x^T + \sigma_n^2 \Sigma^{-1}\right) x \end{aligned}$$

$M = x x^T + \sigma_n^2 \Sigma^{-1}$ $v = M^{-1} y x$.

The machinery at the core of this simplification occurs frequently when dealing with Gaussian densities, and is well worth adding to your toolbox.

We give a lot of details below, though in books and papers, the simplification is often given in a single line. We begin with a bit of rearrangement:

$$\begin{aligned}\mathcal{N}(y\,;\,w^Tx,\sigma_n^2)\mathcal{N}(w\,;\,0,\Sigma_p) &= \frac{1}{\sigma_n\sqrt{2\pi}}\exp\left(-\frac{(y-w^Tx)^2}{2\sigma_n^2}\right)\\ &\quad\times|2\pi\Sigma_p|^{-1/2}\exp\left(-\frac{1}{2}w^T\Sigma^{-1}w\right)\\ &= k\exp\left(-\frac{1}{2}\sigma_n^{-2}\left[(y-w^Tx)^2+\sigma_n^2w^T\Sigma^{-1}w\right]\right),\end{aligned}$$

where $k=\frac{1}{\sigma_n\sqrt{2\pi}}|2\pi\Sigma_p|^{-1/2}$ collects the factors outside the $\exp(\cdot)$. We can now simplify the expression inside the exponential. The goal is to get it into the form:

$$a\,(w-v)^T\,M(w-v)+c,$$

where a and c are just scalar constants, v is a vector independent of w, and M is a symmetric positive definite matrix (spd), independent of w. Once it's in this form, we can write the whole expression as something proportional to a single Gaussian density, rather than the product of two densities. We can get there by a method known as "completing the square" or "completing the quadratic form". We first do some matrix algebra write things in the form $\alpha+w^T\beta+w^THw$:

$$\begin{aligned}&(y-w^Tx)^2+\sigma_n^2w^T\Sigma^{-1}w\\ =\;&\left(y-2yw^Tx+(w^Tx)(x^Tw)\right)+w^T\left(\sigma_n^2\Sigma^{-1}\right)w\\ =\;&y-2yw^Tx+w^T\left(xx^T+\sigma_n^2\Sigma^{-1}\right)w.\end{aligned}\tag{3.1}$$

Note that this is a sum of 3 terms: the first term is a "constant term", independent of w, the second term is linear in w, and the third term is a quadratic form in w. We now expand out the target form that we are going for, and simply equate corresponding parts:

$$(w-v)^T\,M(w-v)+c \;=\; \underbrace{v^TMv+c}_{\text{constant in } w}\; -2v^TMw+w^TMw \tag{3.2}$$

where we've used the symmetry of M to combine the two linear terms. Equating the quadratic terms in (3.1) and (3.2), we get

$$w^TMw=w^T\left(xx^T+\sigma_n^2\Sigma^{-1}\right)w.$$

So we take $M = xx^T + \sigma_n^2\Sigma^{-1}$. (Note that M is indeed spd, since xx^T is positive semidefinite, and Σ^{-1} is spd.) Equating the linear terms, we have

$$\begin{aligned} -2yw^Tx &= -2v^TMw \\ \implies w^T(yx) &= w^T(Mv) \end{aligned}$$

Since M is spd, it is invertible, and we can take $v = M^{-1}yx$. In practice, we often don't need an explicit form for the constant c. In our case, c will eventually just be part of the normalization term for a Gaussian density. For completeness, we give here an explicit form for c by equating the constant terms:

$$\begin{aligned} v^TMv + c &= y \\ \implies c &= y - v^TMv \\ &= y - yx^TM^{-1}MM^{-1}yx \\ &= y - y^2x^TM^{-1}x. \end{aligned}$$

Bringing it all together, we get

$$(y - w^Tx)^2 + \sigma_n^2 w^T\Sigma^{-1}w = (w-v)^T M(w-v) + c,$$

where M, v, and c are as defined above. And so

$$\begin{aligned} \mathcal{N}(y\,;\, w^Tx, \sigma_n^2)\mathcal{N}(w\,;\,0,\Sigma_p) &= k\exp\left(-\frac{1}{2}\sigma_n^{-2}\left[(w-v)^TM(w-v)+c\right]\right) \\ &= k'\exp\left(-\frac{1}{2}(w-v)^T\sigma_n^{-2}M(w-v)\right), \end{aligned}$$

for a new constant $k' = k\exp\left(-\frac{1}{2}\sigma_n^{-2}c\right)$. Bringing it back to our original expression:

$$\begin{aligned} p(w \mid \mathcal{D}) &= \frac{\mathcal{N}(y\,;\, w^Tx, \sigma_n^2)\mathcal{N}(w\,;\,0,\Sigma_p)}{p(y \mid x)} \\ &= k''\exp\left(-\frac{1}{2}(w-v)^T\sigma_n^{-2}M(w-v)\right), \end{aligned}$$

where $k'' = k'/p(y \mid x)$. Note that k'' is still independent of w.

At this point, we claim that the RHS is exactly a multivariate Gaussian density. Because we've been careful to keep track of the explicit expression for

k'', one could verify explicitly that k'' is indeed the appropriate normalization constant for the multivariate Gaussian with variance $\sigma_n^2 M^{-1}$. However, we can avoid this work (and in the future, avoid keeping explicit track of the normalization constants), since we know the following things:

1. $p(w \mid \mathcal{D})$ gives the density for w. So the expression on the RHS must also be a density for w.

2. The RHS is proportional to a multivariate Gaussian density.

Since the RHS is both proportional to a multivariate Gaussian density, and it actually is a density, it must in fact actually be a multivariate Gaussian density. We conclude that

$$p(w \mid \mathcal{D}) = \mathcal{N}(w\,;\, \mu_\pi, \Sigma_\pi),$$

where

$$\begin{aligned} \mu_\pi &= v = M^{-1} y x = y \left(x x^T + \sigma_n^2 \Sigma^{-1}\right)^{-1} x \\ \Sigma_\pi &= \sigma_n^2 M^{-1} = \sigma_n^2 \left(x x^T + \sigma_n^2 \Sigma^{-1}\right)^{-1} = \left(\sigma_n^{-2} x x^T + \Sigma^{-1}\right)^{-1} \end{aligned}$$

3.2 Posterior for Multiple Observations

Above we considered a dataset consisting of a single observation. This can be a realistic scenario in practice: we may get new observations one at a time, and we may want to update our posterior distribution after each observation. We can use the update rules given above. [In homework, we show that updating one data point at a time is equivalent to updating all at once.]

- **Data:** $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
  - Write $y = (y_1, \ldots, y_n)$ and $x = (x_1, \ldots, x_n)$.
- **Design matrix** $X \in \mathbf{R}^{n \times d}$ has input vectors as rows:

$$X = \begin{pmatrix} -x_1- \\ \vdots \\ -x_n- \end{pmatrix}.$$

3.3 (Note: We don't have or need a model for x – we always condition on x, or assume that x is designed.)

We can find that the data likelihood is

$$p(y|X,w) \sim N(X'w, \sigma_n^2 I)$$

and the posterior on the parameters is

$$p(w|X,y) \sim N\left(\bar{w} = \frac{1}{\sigma_n^2} A^{-1} X y, A^{-1}\right)$$

where $A = \sigma_n^{-2} X X' + \Sigma_p^{-1}$. Note this is some combination of the prior and the data covariances. The predictive distribution for a new input point x_* is

$$p(f_*|x_*,X,y) = N\left(\frac{1}{\sigma_n^2} x_*' A^{-1} X y, x_*' A^{-1} x_*\right)$$

3.3.1 Sanity check – noise free case

We should be able to show that in the noise-free case ($\sigma_n^2 = 0$), the marginal distribution of the posterior function of f is degenerate at the training point output value... Say we have two training points x_1 and x_2, and our test point is x_1. Then the posterior mean at x_1 is given in our formulae to have

$$\begin{aligned} \bar{f}_* &= (k_{11}, k_{12}) \begin{pmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{pmatrix}^{-1} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = (1\ 0) \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = y_1 \\ \mathrm{Var}(f_*) &= k_{11} - (k_{11}, k_{12}) \begin{pmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{pmatrix}^{-1} \begin{pmatrix} k_{11} \\ k_{12} \end{pmatrix} = k_{11} - (1\ 0) \begin{pmatrix} k_{11} \\ k_{12} \end{pmatrix} = 0 \end{aligned}$$