Bayesian Linear Regression

David S. Rosenberg

TODO:

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 A Note on Notation

In many texts it's common to denote a random variable with a captial letter, while a particular instantiation of that variable would be denoted with the corresponding lowercase letter. For example: $p(Y = y \mid X = x)$. In our development below, we would simply write $p(y \mid x)$, since it should always be clear in our context what is random and what is not.

We use capital letters to denote matrices (e.g. the design matrix X and the covariance matrix Σ) and lower case letters to denote vectors and scalars.

1.1 Conditional distributions

throughout, everything we write below can be thought of as "conditional on X". if it's a random variable, we can write it down on the right side of the conditional.. but we can also just take it as convention that X is known at every stage, and we can use it in any expression... Thus we mean the same thing by each of the following three expressions:

$$p(\mathcal{D}) = p(y \mid X) = p(y)$$

2 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:

$$\begin{aligned} f(x) &= w^T x \\ \varepsilon &\sim \mathcal{N}(0, \sigma^2) \\ y &= f(x) + \varepsilon. \end{aligned} \tag{2.1}$$

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 $\mathcal{N}(0, \sigma^2)$.
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:

$$Y|x \sim \mathcal{N}(w^T x, \sigma^2). \tag{2.2}$$

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 2.2, 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

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

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 3 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.

3 Maximum Likelihood Estimation

Recall from (2.2) 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

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

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

$$p_w(\mathcal{D}) = \prod_{i=1}^{n} p_w(y_i \mid x_i). \tag{3.1}$$

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 w that has maximum likelihood for the data $\mathcal{D}$. This estimator, know as the **maximum likelihood estimator**, or **MLE**, is $w^* = \arg\max_w p_w(\mathcal{D})$. It is often convenient to express the MLE in terms of the log-likelihood, since it changes the expression in (3.1) from a product into a sum:

$$w^* = \arg\max_{w \in \mathbf{R}^d} \sum_{i=1}^{n} \log p_w(y_i \mid x_i).$$

Let us derive an expression for the MLE w_*. The log-likelihood is

$$\begin{aligned}
\log p_w(\mathcal{D}) &= \sum_{i=1}^{n} \log p_w(y_i \mid x_i) \\
&= \sum_{i=1}^{n} \log\left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2}\right)\right] \\
&= \underbrace{\sum_{i=1}^{n} \log\left[\frac{1}{\sigma\sqrt{2\pi}}\right]}_{\text{independent of } w} + \sum_{i=1}^{n} \left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2}\right)
\end{aligned} \tag{3.2}$$

It is now straightforward[2] to see that we can write

$$w^* = \arg\min_{w \in \mathbf{R}^d} \sum_{i=1}^{n} (y_i - w^T x_i)^2.$$

Hopefully, this is recognizable as the objective function for least squares regression. The take-away message so far is that **maximum likelihood estimation for gaussian linear regression is equivalent to least squares linear regression.**

For completeness, we'll derive a closed form expression for w^*. First, let's rewrite this in matrix notation. Introduce the **design matrix** $X \in \mathbf{R}^{n \times d}$, which has input vectors as rows:

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

and let $y = (y_1, \ldots, y_n)^T$ be the corresponding column vector of responses. Then

$$\begin{aligned} \sum_{i=1}^{n} (y_i - w^T x_i)^2 &= (y - Xw)^T (y - Xw) \\ &= y^T y - 2w^T X^T y + w^T X^T X w. \end{aligned}$$

Since we are minimizing this function over $w \in \mathbf{R}^d$, and $\mathbf{R}^d$ is an open set, the minimum must occur at a critical point. Differentiating with respect to w and equating to 0, we get

$$\begin{aligned} 2X^T X w - 2X^T y &= 0 \\ \iff X^T X w &= X^T y \end{aligned} \tag{3.3}$$

This last expression represents what are often called the **normal equations**[3]. If we assume $X^T X$ is invertible, then a bit of algebra gives the solution as

$$w^* = \left(X^T X\right)^{-1} X^T y.$$

[2] First, note that the first term in the last expression (3.2) is independent of w, and thus we can drop it without changing the maximizer w^*. Similarly, we can drop the factor σ^2 in the second term without affecting the maximizer. Finally, we can flip the sign of the objective function and change the maximization to a minimization, again without affecting w^*

[3] They are called the normal equations because, after rewriting as $X^T (y - Xw) = 0$, we see they express that the residual vector $y - Xw$ is normal to the column space of X.

However, $X^T X$ may not be invertible. For example, if X short and wide ($n < d$ case), or more generally, if X does not have full column rank, then $X^T X$ will not be invertible. This is the **underdetermined** case, in which there are infinitely many equivalent solutions. One can show this with some linear algebra, but this is not (or should not be) an important case for machine learning practice. In the underdetermined case (and in general, unless we have $n \gg d$), we should use **regularized** maximum likelihood, in which case we don't run into this problem.

EXERCISE?

4 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 $w \sim \mathcal{N}(0, \Sigma_p)$. The expressions below give a recipe for generating a dataset of $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ under this model. Note that we assume that $x_1, \ldots, x_n$ are given, and we are generating corresponding random values for $y_1, \ldots, y_n$:

$$\begin{aligned} w &\sim \mathcal{N}(0, \Sigma) \\ f(x_i) &= w^T x_i \text{ for } i = 1, \ldots, n \\ \varepsilon_i &\sim \mathcal{N}(0, \sigma^2) \text{ i.i.d for } i = 1, \ldots, n \\ y_i &= f(x_i) + \varepsilon_i. \end{aligned}$$

We assume that both σ^2 and Σ are known.

We have now written down a full Bayesisan model for our data generating process. Note that we have a fully specified probability distribution for $Y_i \mid x_i$ – there are no "unknown parameters" in the way that w was unknown in the maximum likelihood approach of Section (3). In this Bayesian model, w is an unobserved random variable: mathematically, it has the same status as the ε_i's. In Equations (2.1) and (2.2), we had a collection of candidate probability distributions for $y|x$, one for each value of w.

4.1 Matrix Form

It will be convenient to rewrite this model more compactly, using random vectors. For the data $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, let $y = (y_1, \ldots, y_n)$ and

denote the **design matrix** by $X \in \mathbf{R}^{n \times d}$, which has the input vectors as rows:

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

If we let $f(X) = (f(x_1), \ldots, f(x_n))^T$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$, then we can write

$$\begin{aligned} w &\sim \mathcal{N}(0, \Sigma) \\ f(X) &= Xw \\ \varepsilon &\sim \mathcal{N}(0, \sigma^2 I) \\ y &= f(X) + \varepsilon, \end{aligned}$$

where I denotes the $n \times n$ identity matrix[4]. We can write this in a compact form as

$$\begin{aligned} w &\sim \mathcal{N}(0, \Sigma) \\ y \mid X, w &\sim \mathcal{N}(Xw, \sigma^2 I) \end{aligned}$$

througho

4.2 Posterior

So far, we've defined our Bayesian model. As noted above, this amounts to a specific conditional distribution for $y \mid X$.

$$p(w \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid w)p(w)}{p(\mathcal{D})}.$$

Going to proportionality (important technique!)

We're about to rewrite the expression above as

$$p(w \mid \mathcal{D}) \propto p(\mathcal{D} \mid w)p(w).$$

The $\propto$ is read "is proportional to". This is not a "hand-wavy" expression – it has a precise mathematical meaning. It means that for every each dataset $\mathcal{D}$,

[4] In this form, it's clear that we can generalize this model by replacing $\sigma^2 I$ with a general covariance matrix.

we have a proportionality constant k such that $p(w \mid \mathcal{D}) = kp(\mathcal{D} \mid w)p(w)$. Put another way, there is a function $k(\mathcal{D})$ such that

$$p(w \mid \mathcal{D}) = k(\mathcal{D})p(\mathcal{D} \mid w)p(w) \quad \forall w, \mathcal{D}.$$

So

$$\begin{aligned} p(w \mid \mathcal{D}) &\propto p(\mathcal{D} \mid w)p(w) \\ &= p(y \mid X, w)p(w) \\ &= \mathcal{N}\left(y\,;\, Xw, \sigma^2 I\right)\mathcal{N}(w\,;\, 0, \Sigma) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}(y - Xw)^T(y - Xw)\right)\exp\left(-\frac{1}{2}w^T\Sigma^{-1}w\right) \\ &= \exp\left(-\frac{1}{2}\left[\frac{1}{\sigma^2}(y - Xw)^T(y - Xw) + w^T\Sigma^{-1}w\right]\right). \end{aligned}$$

Extracting out the piece in the exponent,

$$\begin{aligned} &\frac{1}{\sigma^2}(y - Xw)^T(y - Xw) + w^T\Sigma^{-1}w \\ = &\frac{1}{\sigma^2}\left(w^TX^TXw - 2w^TX^Ty + y^Ty\right) + w^T\Sigma^{-1}w \\ = &w^T\left(\frac{1}{\sigma^2}X^TX + \Sigma^{-1}\right)w - 2\left(\frac{1}{\sigma^2}\right)y^TXw. \end{aligned}$$

To simplify our expressions, let's take $M = \frac{1}{\sigma^2}X^TX + \Sigma^{-1}$ and $b = \left(\frac{1}{\sigma^2}\right)X^Ty$. The expression them becomes $w^TMw - 2b^Tw$. We can now "complete the quadratic form" by applying following identity

$$w^TMw - 2b^Tw = \left(w - M^{-1}b\right)^T M(w - M^{-1}b) - b^TM^{-1}b,$$

which is easily verified by expanding the quadratic form on the RHS. We call it "completing the quadratic form" because while the LHS has both quadratic and linear terms involving w, while on the RHS w only appears in a quadratic term. (For a slower introduction to this technique, see the notes on Com-

pleting the Quadratic Form.) Putting it together, we get

$$\begin{aligned} p(w \mid \mathcal{D}) &\propto \exp\left(-\frac{1}{2}\left[\left(w - M^{-1}b\right)^T M(w - M^{-1}b) - b^T M^{-1}b\right]\right) \\ &= \exp\left(-\frac{1}{2}\left[\left(w - M^{-1}b\right)^T M(w - M^{-1}b)\right]\right)\exp\left(-\frac{1}{2}\left[-b^T M^{-1}b\right]\right) \\ &\propto \exp\left(-\frac{1}{2}\left[\left(w - M^{-1}b\right)^T M(w - M^{-1}b)\right]\right). \end{aligned}$$

Now recall that a Gaussian density in w is given by

$$\mathcal{N}(w; \mu, \Sigma) = |2\pi\Sigma|^{-1/2}\exp\left(-\frac{1}{2}(w-\mu)^T\Sigma^{-1}(w-\mu)\right).$$

So

$$p(w \mid \mathcal{D}) \propto \mathcal{N}\left(w; M^{-1}b, M^{-1}\right) \tag{4.1}$$

Since the LHS and RHS of (4.1) are both densities in w and are proportional, they must actually be equal[5]:

$$p(w \mid \mathcal{D}) = \mathcal{N}\left(w; M^{-1}b, M^{-1}\right),$$

where $M = \frac{1}{\sigma^2}X^TX + \Sigma^{-1}$ and $b = \left(\frac{1}{\sigma^2}\right)X^Ty$.

Note that the posterior mean is

$$\begin{aligned} M^{-1}b &= \left(\frac{1}{\sigma^2}X^TX + \Sigma^{-1}\right)^{-1}\frac{1}{\sigma^2}X^Ty \\ &= \left(X^TX + \sigma^2\Sigma^{-1}\right)^{-1}X^Ty, \end{aligned}$$

which should look familiar from our study of ridge regression. Indeed, if the prior covariance matrix is taken to be $\Sigma = \frac{\sigma^2}{\lambda}I$, then the posterior mean is

$$\left(X^TX + \lambda I\right)^{-1}X^Ty,$$

which is exactly the ridge regression estimate for w.

To make things look prettier, people often specify the gaussian prior in terms of the **precision matrix**, which is the inverse of the covariance matrix. That is $\Lambda = \Sigma^{-1}$. Then the posterior mean looks like

$$\left(X^TX + \sigma^2\Lambda\right)^{-1}X^Ty.$$

[5] See notes on proportionality for a bit more discussion of this idea.

The precision matrix of a Gaussian distribution has some other interesting properties as well (see ...).

which is of course the ridge regression solution. S

Thus we have derived the posterior distribution for the unknown parameter w conditioned on the data $\mathcal{D}$. We write this result in the theorem below, purely in terms of probability distributions, without mentioning "priors" or "posteriors".

Theorem 1. *Given a fixed* ***design matrix*** $X \in \mathbf{R}^{m \times n}$*, and a random vector* $y = Xw + \varepsilon$*, where* $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ *and* $w \sim \mathcal{N}(0, \Sigma)$*, the conditional distribution of* $w \mid y$ *is* $\mathcal{N}(m, V)$*, where*

$$\begin{aligned} m &= \left(X^T X + \sigma^2 \Sigma^{-1}\right)^{-1} X^T y \\ V &= \sigma^2 \left(X^T X + \sigma^2 \Sigma^{-1}\right)^{-1}. \end{aligned}$$

4.2.1 Predictive Distributions

In machine learning contexts, our ultimate goal is typically prediction, rather than parameter estimation. That is, our primary objective is typically to predict the y corresponding to a new x, rather than to estimate w. Predictive distributions are straightforward to calculate. Before seeing any data, the predictive distribution for x is simply

$$\begin{aligned} p(y \mid x) &= \int_w p(y \mid w) p(w)\, dw \\ &= \end{aligned}$$

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

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

4.3 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:

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

- The data likelihood is

$$\begin{aligned} p(\mathcal{D} \mid w) &= p(y \mid x, w) \\ &= p(\\ &\quad \mathcal{N}(y\,;\,) \\ &\quad \prod_{i=1}^{n} p(y_i \mid x_i, w) \\ &= \end{aligned}$$

  The posterior distribution is

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

4.4 (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

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