

# ELEN 4903: Machine Learning

## Week 3, Lecture 1, 1/31/2018

Prof. John Paisley

Department of Electrical Engineering  
& Data Science Institute

Columbia University

# BAYESIAN LINEAR REGRESSION

## Model

Have vector  $y \in \mathbb{R}^n$  and covariates matrix  $X \in \mathbb{R}^{n \times d}$ . The  $i$ th row of  $y$  and  $X$  correspond to the  $i$ th observation  $(y_i, x_i)$ .

In a Bayesian setting, we model this data as:

$$\mathbf{Likelihood} : y \sim N(Xw, \sigma^2 I)$$

$$\mathbf{Prior} : w \sim N(0, \lambda^{-1} I)$$

The unknown model variable is  $w \in \mathbb{R}^d$ .

- ▶ The “likelihood model” says how well the observed data agrees with  $w$ .
- ▶ The “model prior” is our prior belief (or constraints) on  $w$ .

This is called Bayesian linear regression because we have defined a prior on the unknown parameter and will try to learn its posterior.

# REVIEW: MAXIMUM A POSTERIORI INFERENCE

## MAP solution

MAP inference returns the maximum of the log joint likelihood.

$$\textbf{Joint Likelihood} : \quad p(y, w|X) = p(y|w, X)p(w)$$

Using Bayes rule, we see that this point also maximizes the *posterior* of  $w$ .

$$\begin{aligned} w_{\text{MAP}} &= \arg \max_w \ln p(w|y, X) \\ &= \arg \max_w \ln p(y|w, X) + \ln p(w) - \ln p(y|X) \\ &= \arg \max_w -\frac{1}{2\sigma^2}(y - Xw)^T(y - Xw) - \frac{\lambda}{2}w^T w + \text{const.} \end{aligned}$$

We saw that this solution for  $w_{\text{MAP}}$  is the same as for ridge regression:

$$w_{\text{MAP}} = (\lambda\sigma^2 I + X^T X)^{-1} X^T y \quad \Leftrightarrow \quad w_{\text{RR}}$$

# POINT ESTIMATES VS BAYESIAN INFERENCE

## Point estimates

$w_{\text{MAP}}$  and  $w_{\text{ML}}$  are referred to as *point estimates* of the model parameters.

They find a specific value (point) of the vector  $w$  that maximizes an objective function — the posterior (MAP) or likelihood (ML).

- ▶ **ML:** Only considers the data model:  $p(y|w, X)$ .
- ▶ **MAP:** Takes into account model prior:  $p(y, w|X) = p(y|w, X)p(w)$ .

## Bayesian inference

Bayesian inference goes one step further by characterizing uncertainty about the values in  $w$  using Bayes rule.

# BAYES RULE AND LINEAR REGRESSION

## Posterior calculation

Since  $w$  is a continuous-valued random variable in  $\mathbb{R}^d$ , Bayes rule says that the *posterior* distribution of  $w$  given  $y$  and  $X$  is

$$p(w|y, X) = \frac{p(y|w, X)p(w)}{\int_{\mathbb{R}^d} p(y|w, X)p(w) dw}$$

That is, we get an updated distribution on  $w$  through the transition

prior  $\rightarrow$  likelihood  $\rightarrow$  posterior

**Quote:** “The posterior of \_\_ is proportional to the likelihood times the prior.”

# FULLY BAYESIAN INFERENCE

## Bayesian linear regression

In this case, we can update the posterior distribution  $p(w|y, X)$  analytically.

We work with the proportionality first:

$$\begin{aligned} p(w|y, X) &\propto p(y|w, X)p(w) \\ &\propto \left[ e^{-\frac{1}{2\sigma^2}(y-Xw)^T(y-Xw)} \right] \left[ e^{-\frac{\lambda}{2}w^Tw} \right] \\ &\propto e^{-\frac{1}{2}\{w^T(\lambda I + \sigma^{-2}X^TX)w - 2\sigma^{-2}w^TX^Ty\}} \end{aligned}$$

The  $\propto$  sign lets us multiply and divide this by anything *as long as it doesn't contain  $w$* . We've done this twice above. Therefore the 2nd line  $\neq$  3rd line.

# BAYESIAN INFERENCE FOR LINEAR REGRESSION

We need to normalize:

$$p(w|y, X) \propto e^{-\frac{1}{2}\{w^T(\lambda I + \sigma^{-2}X^T X)w - 2\sigma^{-2}w^T X^T y\}}$$

There are two key terms in the exponent:

$$\underbrace{w^T(\lambda I + \sigma^{-2}X^T X)w}_{\text{quadratic in } w} - \underbrace{2w^T X^T y/\sigma^2}_{\text{linear in } w}$$

We can conclude that  $p(w|y, X)$  is Gaussian. Why?

1. We can multiply and divide by anything not involving  $w$ .
2. A Gaussian has  $(w - \mu)^T \Sigma^{-1} (w - \mu)$  in the exponent.
3. We can “complete the square” by adding terms not involving  $w$ .

# BAYESIAN INFERENCE FOR LINEAR REGRESSION

**Compare:** In other words, a Gaussian looks like:

$$p(w|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(w^T \Sigma^{-1} w - 2w^T \Sigma^{-1} \mu + \mu^T \Sigma^{-1} \mu)}$$

and we've shown for some setting of  $Z$  that

$$p(w|y, X) = \frac{1}{Z} e^{-\frac{1}{2}(w^T (\lambda I + \sigma^{-2} X^T X) w - 2w^T X^T y / \sigma^2)}$$

**Conclude:** What happens if in the above Gaussian we define:

$$\Sigma^{-1} = (\lambda I + \sigma^{-2} X^T X), \quad \Sigma^{-1} \mu = X^T y / \sigma^2 ?$$

Using these specific values of  $\mu$  and  $\Sigma$  we only need to set

$$Z = (2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}} e^{\frac{1}{2} \mu^T \Sigma^{-1} \mu}$$



# BAYESIAN INFERENCE FOR LINEAR REGRESSION

## The posterior distribution

Therefore, the posterior distribution of  $w$  is:

$$\begin{aligned}p(w|y, X) &= N(w|\mu, \Sigma), \\ \Sigma &= (\lambda I + \sigma^{-2}X^T X)^{-1}, \\ \mu &= (\lambda \sigma^2 I + X^T X)^{-1} X^T y \quad \Leftarrow \quad w_{\text{MAP}}\end{aligned}$$

Things to notice:

- ▶  $\mu = w_{\text{MAP}}$  after a redefinition of the regularization parameter  $\lambda$ .
- ▶  $\Sigma$  captures uncertainty about  $w$ , like  $\text{Var}[w_{\text{LS}}]$  and  $\text{Var}[w_{\text{RR}}]$  did before.
- ▶ However, now we have a full probability distribution on  $w$ .

# USES OF THE POSTERIOR DISTRIBUTION

## Understanding $w$

We saw how we could calculate the variance of  $w_{LS}$  and  $w_{RR}$ . Now we have an entire distribution. Some questions we can ask are:

**Q:** Is  $w_i > 0$  or  $w_i < 0$ ? Can we confidently say  $w_i \neq 0$ ?

**A:** Use the *marginal posterior distribution*:  $w_i \sim N(\mu_i, \Sigma_{ii})$ .

**Q:** How do  $w_i$  and  $w_j$  relate?

**A:** Use their joint marginal posterior distribution:

$$\begin{bmatrix} w_i \\ w_j \end{bmatrix} \sim N \left( \begin{bmatrix} \mu_i \\ \mu_j \end{bmatrix}, \begin{bmatrix} \Sigma_{ii} & \Sigma_{ij} \\ \Sigma_{ji} & \Sigma_{jj} \end{bmatrix} \right)$$

## Predicting new data

The posterior  $p(w|y, X)$  is perhaps most useful for predicting new data.

# PREDICTING NEW DATA

# PREDICTING NEW DATA

**Recall:** For a new pair  $(x_0, y_0)$  with  $x_0$  measured and  $y_0$  unknown, we can predict  $y_0$  using  $x_0$  and the LS or RR (i.e., ML or MAP) solutions:

$$y_0 \approx x_0^T w_{\text{LS}} \quad \text{or} \quad y_0 \approx x_0^T w_{\text{RR}}$$

With Bayes rule, we can make a *probabilistic* statement about  $y_0$ :

$$\begin{aligned} p(y_0|x_0, y, X) &= \int_{\mathbb{R}^d} p(y_0, w|x_0, y, X) dw \\ &= \int_{\mathbb{R}^d} p(y_0|w, x_0, y, X) p(w|x_0, y, X) dw \end{aligned}$$

Notice that *conditional independence* lets us write

$$p(y_0|w, x_0, y, X) = \underbrace{p(y_0|w, x_0)}_{\text{likelihood}} \quad \text{and} \quad p(w|x_0, y, X) = \underbrace{p(w|y, X)}_{\text{posterior}}$$

# PREDICTING NEW DATA

## Predictive distribution (intuition)

This is called the *predictive distribution*:

$$p(y_0|x_0, y, X) = \int_{\mathbb{R}^d} \underbrace{p(y_0|x_0, w)}_{\text{likelihood}} \underbrace{p(w|y, X)}_{\text{posterior}} dw$$

Intuitively:

1. Evaluate the likelihood of a value  $y_0$  given  $x_0$  for a particular  $w$ .
2. Weight that likelihood by our current belief about  $w$  given data  $(y, X)$ .
3. Then sum (integrate) over all possible values of  $w$ .

# PREDICTING NEW DATA

We know from the model and Bayes rule that

$$\text{Model: } p(y_0|x_0, w) = N(y_0|x_0^T w, \sigma^2),$$

$$\text{Bayes rule: } p(w|y, X) = N(w|\mu, \Sigma).$$

With  $\mu$  and  $\Sigma$  calculated on a previous slide.

The predictive distribution can be calculated exactly with these distributions.  
Again we get a Gaussian distribution:

$$p(y_0|x_0, y, X) = N(y_0|\mu_0, \sigma_0^2),$$

$$\mu_0 = x_0^T \mu,$$

$$\sigma_0^2 = \sigma^2 + x_0^T \Sigma x_0.$$

Notice that the expected value is the MAP prediction since  $\mu_0 = x_0^T w_{\text{MAP}}$ , but we now quantify our confidence in this prediction with the variance  $\sigma_0^2$ .

# ACTIVE LEARNING

## PRIOR $\rightarrow$ POSTERIOR $\rightarrow$ PRIOR

Bayesian learning is naturally thought of as a sequential process. That is, the posterior after seeing some data becomes the prior for the next data.

Let  $y$  and  $X$  be “old data” and  $y_0$  and  $x_0$  be some “new data”. By Bayes rule

$$p(w|y_0, x_0, y, X) \propto p(y_0|w, x_0)p(w|y, X).$$

The posterior after  $(y, X)$  has become the prior for  $(y_0, x_0)$ .

Simple modifications can be made sequentially in this case:

$$p(w|y_0, x_0, y, X) = N(w|\mu, \Sigma),$$

$$\Sigma = (\lambda I + \sigma^{-2}(x_0 x_0^T + \sum_{i=1}^n x_i x_i^T))^{-1},$$

$$\mu = (\lambda \sigma^2 I + (x_0 x_0^T + \sum_{i=1}^n x_i x_i^T))^{-1}(x_0 y_0 + \sum_{i=1}^n x_i y_i).$$



# INTELLIGENT LEARNING

Notice we could also have written

$$p(w|y_0, x_0, y, X) \propto p(y_0, y|w, X, x_0)p(w)$$

but often we want to use the sequential aspect of inference to help us learn.

Learning  $w$  and making predictions for new  $y_0$  is a two-step procedure:

- ▶ Form the predictive distribution  $p(y_0|x_0, y, X)$ .
- ▶ Update the posterior distribution  $p(w|y, X, y_0, x_0)$ .

**Question:** Can we learn  $p(w|y, X)$  intelligently?

That is, if we're in the situation where we can pick which  $y_i$  to measure with knowledge of  $\mathcal{D} = \{x_1, \dots, x_n\}$ , can we come up with a good strategy?

## An “active learning” strategy

Imagine we already have a measured dataset  $(y, X)$  and posterior  $p(w|y, X)$ . We can construct the predictive distribution for every remaining  $x_0 \in \mathcal{D}$ .

$$\begin{aligned} p(y_0|x_0, y, X) &= N(y_0|\mu_0, \sigma_0^2), \\ \mu_0 &= x_0^T \mu, \\ \sigma_0^2 &= \sigma^2 + x_0^T \Sigma x_0. \end{aligned}$$

For each  $x_0$ ,  $\sigma_0^2$  tells how confident we are. This suggests the following:

1. Form predictive distribution  $p(y_0|x_0, y, X)$  for all unmeasured  $x_0 \in \mathcal{D}$
2. Pick the  $x_0$  for which  $\sigma_0^2$  is largest and measure  $y_0$
3. Update the posterior  $p(w|y, X)$  where  $y \leftarrow (y, y_0)$  and  $X \leftarrow (X, x_0)$
4. Return to #1 using the updated posterior

## Entropy (i.e., uncertainty) minimization

When devising a procedure such as this one, it's useful to know what *objective function* is being optimized in the process.

We introduce the concept of the *entropy* of a distribution. Let  $p(z)$  be a continuous distribution, then its (differential) entropy is:

$$\mathcal{H}(p) = - \int p(z) \ln p(z) dz.$$

This is a measure of the spread of the distribution. More positive values correspond to a more “uncertain” distribution (larger variance).

The entropy of a multivariate Gaussian is

$$\mathcal{H}(N(w|\mu, \Sigma)) = \frac{1}{2} \ln \left( (2\pi e)^d |\Sigma| \right).$$

# ACTIVE LEARNING

The entropy of a Gaussian changes with its covariance matrix. With sequential Bayesian learning, the covariance transitions from

$$\text{Prior : } (\lambda I + \sigma^{-2} X^T X)^{-1} \equiv \Sigma$$

$\Downarrow$

$$\text{Posterior : } (\lambda I + \sigma^{-2}(x_0 x_0^T + X^T X))^{-1} \equiv (\Sigma^{-1} + \sigma^{-2} x_0 x_0^T)^{-1}$$

Using a “rank-one update” property of the determinant, we can show that the entropy of the prior  $\mathcal{H}_{\text{prior}}$  relates to the entropy of the posterior  $\mathcal{H}_{\text{post}}$  as:

$$\mathcal{H}_{\text{post}} = \mathcal{H}_{\text{prior}} - \frac{d}{2} \ln(1 + \sigma^{-2} x_0^T \Sigma x_0)$$

Therefore, the  $x_0$  that minimizes  $\mathcal{H}_{\text{post}}$  also maximizes  $\sigma^2 + x_0^T \Sigma x_0$ . We are minimizing  $\mathcal{H}$  myopically, so this is called a “greedy algorithm”.

# MODEL SELECTION

# SELECTING $\lambda$

We've discussed  $\lambda$  as a “nuisance” parameter that can impact performance.

Bayes rule gives a principled way to do this via *evidence maximization*:

$$p(w|y, X, \lambda) = \underbrace{p(y|w, X)}_{\text{likelihood}} \underbrace{p(w|\lambda)}_{\text{prior}} / \underbrace{p(y|X, \lambda)}_{\text{evidence}}.$$

The “evidence” gives the likelihood of the data with  $w$  integrated out. It's a measure of how good our model and parameter assumptions are.

# SELECTING $\lambda$

If we want to set  $\lambda$ , we can also do it by maximizing the evidence.<sup>1</sup>

$$\hat{\lambda} = \arg \max_{\lambda} \ln p(y|X, \lambda).$$

We notice that this looks exactly like maximum likelihood, and it is:

**Type-I ML:** Maximize the likelihood over the “main parameter” ( $w$ ).

**Type-II ML:** Integrate out “main parameter” ( $w$ ) and maximize over the “hyperparameter” ( $\lambda$ ). Also called *empirical Bayes*.

The difference is only in their perspective.

This approach requires us to solve this integral, but we often can't for more complex models. Cross-validation is an alternative that's always available.

---

<sup>1</sup>We can show that the distribution of  $y$  is  $p(y|X, \lambda) = N(y|0, \sigma^2 I + \lambda^{-1} X X^T)$ . This would require an algorithm to maximize over  $\lambda$ . The key point here is the general technique.