COMS 4721: Machine Learning for Data Science

Lecture 5, 1/31/2017

Prof. John Paisley

Department of Electrical Engineering
& Data Science Institute

Columbia University

Model

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

Assume that y is generated from a Gaussian with mean equal to xw and covariance λ^{-1}.

In a Bayesian setting, we model this data as:

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

→ given setting w

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

unknown

place a prior distribution on the vector w, which is zero mean Gaussian and covariance λ^{-1}.

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, we're going to define in advance.

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

REVIEW: MAXIMUM A POSTERIORI INFERENCE

We saw last time instead of finding the full posterior, we could find the map solution. By finding the value of the vector W that maximizes the posterior.

MAP solution

When you pick Gaussian prior & likelihood MAP equivalent to ridge.

MAP inference returns the maximum of the log joint likelihood.

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

most probable acc. to posterior distribution.

Using Bayes rule 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_{MAP} is the same as for ridge regression:

corresponds to where we have made a redefinition of the regularization parameter.

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

prior & noise combination

POINT ESTIMATES VS BAYESIAN INFERENCE

The difference of point estimates for model variables in Bayesian inference.

we find a specific value for an unknown model parameter acc. to some inference procedure.

Point estimates

w_{MAP} and w_{ML} are referred to as *point estimates* of the model parameters. (only 1 value)

because they return specific values of the unknown model variables.

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

maximizes this term.

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

Bayesian inference

difference

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

Instead of returning a pt. estimate is going to return probability distribution on the unknown model variable.

Posterior calculation

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

likelihood

normalizing constant.

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

prior

Integral of the numerator over all possible values w can take.

can't solve for complex integrals

sum in continuous space

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

Sequence of bayes rule

$$\text{prior} \quad \rightarrow \quad \text{likelihood} \quad \rightarrow \quad \text{posterior}$$

defined on slide 1

Quote: "The posterior of __ is proportional to the likelihood times the prior."

something

(how we can do fully Bayesian inference with linear regression problem.)

Bayesian linear regression

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

We work with the proportionality first:

both functions of w

eliminated any function that doesn't involve w.

constant before

zero mean Gaussian.

constant observed.

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

$e^{-\frac{1}{2}y^Ty}$ removed

cancel when we normalize.

The $\propto$ sign lets us multiply and divide this by anything *as long as it doesn't contain w*. We've done this in two lines above.

* The posterior distribution of our regression coefficient vector w, given the data is proportional to the likelihood of the responses given w and given the covariance

× times the prior of the regression coefficient vector w.

(to make it an equality)

We need to normalize: (divide this function by its integral over all values of w in $\mathbb{R}^d$.)

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

There are two key terms in the exponent:

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

[complete square & result is gaussian to get this form]

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.

In order to solve Bayes rule we are allowed to multiply & this term* by anything that doesn't involve w. So we can multiply by something/divide by something, not involving w, such that we know that the result integrates to 1.

BAYESIAN INFERENCE FOR LINEAR REGRESSION

$(w-\mu)^T\Sigma^{-1}(w-\mu)$

Compare: In other words, a Gaussian looks like:

$\rightarrow \mu^T\Sigma^{-1}\Sigma\Sigma^{-1}\mu$

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

just multiplying Gaussian terms.

and we've shown for some setting of Z that

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

find value Z such that this function integrates to 1.

*

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

**

$$\Sigma^{-1} = (\lambda I + \sigma^{-2}X^TX), \qquad \Sigma^{-1}\mu = X^Ty/\sigma^2\,?$$

***

Using these specific values of μ and Σ we only need to set

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

*
This just a quick way we calculate the posterior as being Gaussian, without actually ever having to try to solve this integral overw. We can say that, for z, plug in this and then solve, to find that we can simplify the result to be Gaussian with covariance equal to the inverse of this, and the inverse covariance times equal to this.
**
***

what does this mean?

The posterior distribution

Therefore, the posterior distribution of w is:

In other words, posterior of w given data y and x, is a Gaussian with mean equal to this and covariance equal to this.

(multivariate Gaussian)

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

$$\Sigma = (\lambda I + \sigma^{-2}X^TX)^{-1},$$

$$\mu = (\lambda\sigma^2 I + X^TX)^{-1}X^Ty \Leftarrow w_{\text{MAP}}$$

The sol^n for mean of posterior distribution is equal to w_{MAP} sol^n, which we know is also equal to ridge regression sol^n.

Things to notice:

But now we have

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

what is the difference?

In a way, the variance of ridge regression & least square sol^n. also captured some uncertainty of w, but now we actually have a functional distribution that we can work with. So we can give densities and calculate probabilities that we couldn't do when we calculated these variances.

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:

uni-variate Gaussian

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

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

diagonal element of covariance matrix.

depends upon Σ_{ii}

Q: How do w_i and w_j relate?

A: Use their joint marginal posterior distribution:

$$\left[\begin{array}{c} w_i \\ w_j \end{array}\right] \sim N\left(\left[\begin{array}{c} \mu_i \\ \mu_j \end{array}\right], \left[\begin{array}{cc} \Sigma_{ii} & \Sigma_{ij} \\ \Sigma_{ji} & \Sigma_{jj} \end{array}\right]\right)$$

← calculations in Bishop

submatrix of Σ

Predicting new data

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

* By increasing the x_i dimension of x we increase the expected output, whereas if w_i is negative by increasing the i^{th} dimension of x, we decrease the expected output. This tells us how x relates to y. (Imp. for practical purposes).

PREDICTING NEW DATA

using Bayesian
linear regression.

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) outputs:

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

In both cases, we have a point estimate of w and so we give a point estimate of our prediction, with no uncertainity about what we think y_0 is.

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

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

[Marginal prob. over the unknown coefficient vector.]

previously observed data

(integrating uncertainity abt. w)

Factorize the joint distribution into the product

Notice that ***conditional independence*** lets us write

because of our 3 assumption

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

independent of any other data given w & x_0.

no info. abt w.

we're predicting new data given old data.

Predictive distribution (intuition)

This is called the *predictive distribution*: Marginal over the likelihood of new data given model variables.

Posterior distribution of those model variables given the old data.

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

Intuitively, we evaluate the likelihood of a new y_0 for a particular w and observed x_0, and weight it by our current belief about w given data (y, X).

that model variable

We then sum (integrate) over all possible values of w.

weighted models

And we're essentially removing all of the uncertainity of w in forming our predictive distribution.

Predicting new data

We know from the model and Bayes rule that

Model of new data given the model variable.

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

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

Posterior distribution of the regression coefficient vector given y & x.

With μ and Σ calculated on a previous slide. (Pg 8 and 10)

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

(After computing the integral) →

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

posterior mean, map sol^n

confidence in prediction of w.

noise from likelihood.

*

Notice that the expected value is the MAP prediction since $\mu = x_0^T w_{\text{MAP}}$, but we now quantify our confidence in this prediction with the variance σ_0^2.

* We showed the mean of posterior is equal to the map sol^n. And so the mean of our prediction distribution is the map prediction, the point estimate under map. But now we have this additional variance, which is used to give us a sense of how confident we are in what w is
⇒ Which we couldn't have before because we used a point estimate to solve for it, whereas now we have the posterior distribution of w. And so our uncertainity of w is contained in this distribution and we can then use that uncertainity to propagate it forward to the predictive distribution.

ACTIVE LEARNING

How predictive distribution has more uses than simply making predictions on data.

PRIOR → POSTERIOR → PRIOR

a sequential process where we have a prior belief of a model variable. We get data. And through the data, we then calculate our posterior belief of the model variable. And that posterior becomes the prior for the next data.

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

full posterior given all of the data.

posterior of model given all of old data.

So we use the posterior in the prediction distribution to help us, gives us a prior distribution of what we think the model variable be.

$$p(w|y_0, x_0, y, X) \propto \underbrace{p(y_0|w, x_0)}_{\text{likelihood of new data}} p(w|y, X).$$

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

doesn't this give a lot of importance to just a new sample.

Simple modifications can be made sequentially:

posterior for linear regression.

next observation written out separately

$$\begin{aligned} p(w|y_0, x_0, y, X) &= N(w|\mu, \Sigma), \\ \Sigma &= (\lambda I + \sigma^{-2}(x_0x_0^T + \textstyle\sum_{i=1}^n x_i x_i^T))^{-1}, \\ \mu &= (\lambda\sigma^2 I + (x_0x_0^T + \textstyle\sum_{i=1}^n x_i x_i^T)^{-1}(x_0y_0 + \textstyle\sum_{i=1}^n x_i y_i). \end{aligned}$$

treat as a single matrix we update this way.

?

a sufficient statistic we just keep.

cumulative sum of outer products

cumulative sum of scalar products

treat as a single vector that we then update this way.

Active learning - a general technique for learning model efficiently.

* Bayesian modelling → a sequential process where we see a sequence of data And given everything that we know up until a certian point, we calculate our posterior belief of the model variables. We use that to then make a prediction for the next observation. And after seeing the next observation we update our posterior belief. And the idea is to do so efficiently.

(Idea of active learning: is there a sequence that we can follow to somehow observe data that is going to give us a lot of information towards the end goal of calculating our posterior on the model variables)

** could also make likelihood over the last observation, and the prior in that context becomes the posterior given all the previous observations.

INTELLIGENT LEARNING (How we can use this sequential Bayesian learning to learn the model posterior more efficiently.)

Of course, 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 $p(y_0|x_0, y, X)$. corresponds to the response of new covariate vector that we measure, given all of the old data.
- Update the posterior distribution $p(w|y, X, y_0, x_0)$.

(Using that measured response we update the posterior as shown in previous slide, by ① a rank 1 update ② or adding a vector to a sufficient statistic.)

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

How we can efficiently learn the posterior when we get choose a sequence of measurements observations we have.

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

We want a way of sequentially picking these responses so that we're going get a lot of information about our posterior of w.

* quadratic product between x_0 and posterior covariance, where the covariance is from the Gaussian posterior of w

→ what does this mean? why is it here?

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 $x_0 \in \mathcal{D}$.

? how does it change with x_0 & why?

(these are all the x_0 in our dataset for which we have a corresponding response)

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

μ → Posterior mean

tells about confidence we are in our prediction. So, for different x_0 we're going to have different measurements of how confident we are in our prediction of y_0 in a sense.

For each x_0, σ_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 σ_0^2 is largest and measure y_0 (uncertainty is greatest.)
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

↑ augmented data

Every step we're choosing to measure the response that corresponds to a covariate vector that we're the least certain about our prediction.

Now, question what is it we're trying to optimise when we do this?

Differential entropy because it

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:

This a function of some continous random variable z.

*

as it bigger, you get uncertain

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

(uncertainty of z)

↑ var ↑ entropy

This is a measure of the spread of the distribution. Larger values correspond to a more "uncertain" distribution (more variance). *

The entropy of a multivariate Gaussian is

dimensionality of vector w.

covariance

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

[doesn't depend on μ)

↳ how spread out Gaussian is. [↑ more spread]

*

So a larger value of the differential entropy
in a sense corresponds to a distribution that has a smaller variance. how?
It's a distribution where we're more confident
in the region of values that z can take.
Whereas when the differential entropy becomes more and
more negative, then we're less uncertain about z.
And in a sense,
we can think of this distribution as having a very large variance.

Distributions that have larger variance will have larger differential entropy. As the variance gets smaller and smaller, the differential entropy will become more & more -ve. To point where variance goes to a point estimate, so you have no uncertainity the differential entropy goes to negative infinity.

ACTIVE LEARNING

(We can use the differential entropy to show that the active learning procedure, is a greedy algorithm whereby we are picking the covariance to measure their associated responses that give us the most information about the posterior belief on w.) And information is measured by the differential entropy of our posterior. *

The entropy of a Gaussian changes with its covariance matrix. With sequential Bayesian learning, the <u>covariance transition</u>s from

** posterior of previous observation

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

updation of

$$\Downarrow \text{ after measuring } x_0$$

$$\text{Posterior}: \quad (\lambda I + \sigma^{-2}(x_0 x_0^T + X^T X))^{-1} \equiv (\Sigma^{-1} + \underbrace{\sigma^{-2} x_0 x_0^T}_{\text{only depends on our next measurement}})^{-1}$$

doesn't depend upon y. [So we don't actually need to know what the measurements are in order to calculate the posterior uncertainty of w.]

Using a rank-one update property of the determinant, the entropy of the prior $\mathcal{H}_{\text{prior}}$ is related to the entropy of the posterior $\mathcal{H}_{\text{post}}$ as follows:

pick x_0 that minimizes this & thus maximizes this

covariate vector we chose to measure

prior covariance, calculated using all of measurements we have up til this point.

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

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".

same rule we discussed before, we pick x_0 to the vector for which our predictive uncertainity is the greatest. Viewing it from the perspective of minimizing our posterior uncertainity. ***

* So the less uncertain we are about w, the more and more negative our differential entropy of w will be.

** Toshov: The previous algorithm is a way of picking a point to measure. That's going to minimize the posterior differential entropy among all of the options that we.

*** Trying to pick the maximum of our predictive uncertainity viewed from the perspective of minimizing of our posterior uncertainity results in identically the same rule.

So in this sense, we can see that we have a greedy algorithm because we're only picking the next point to minimize in a greedy way.

Our objective function where the objective is the posterior uncertainity of w.

Model selection

(in context of Bayesian inference)

Prior distribution on w is a zero mean Gaussian where the precision matrix of the inverse of the covariance is λI

We've discussed λ as a "nuisance" parameter that can impact performance.

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

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

now also conditioning on prior.

→ distribution of the data given the hyper parameter λ with w integrated out.

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 λ

If we want to set λ, we can also do it by maximizing the evidence.

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

marginal likelihood of data where we have integrated out of unknown model variables.

We can show that the distribution of y is $p(y|X, \lambda) = N(y|0, \sigma^2 I + \lambda^{-1} X^T X)$. In this case, that's w (not derive it)
This requires an algorithm to maximize over λ. for this distribution

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

Type-I ML: Maximize the likelihood over the "main parameter" (w).

By maximizing the marginal likelihood / doing evidence maximization, where we integrate out the main parameter.

Type-II ML: Integrate out "main parameter" (w) and maximize over the "hyperparameter" (λ). Also called *empirical Bayes*.

The difference is only in their perspective.

This approach requires that we can solve this integral, but often we can't for more complex models. Cross-validation is the method that always works.

In the 1^{st} case, we maximized over what we called the main parameter. So there was no distribution on that parameter and we didn't think of that parameter as being a marginal where we have integrated out everything else.

$p(y|x,w)$ over w.

Now we're maximizing the evidence which is marginal likelihood where we have integrated out the model parameters.