ColumbiaX: Machine Learning

Lecture 3

Prof. John Paisley

Department of Electrical Engineering
& Data Science Institute

Columbia University

Regression: Problem Definition

Data

Measured pairs (x, y), where $x \in \mathbb{R}^{d+1}$ (input) and $y \in \mathbb{R}$ (output)

Goal

Find a function $f : \mathbb{R}^{d+1} \to \mathbb{R}$ such that $y \approx f(x; w)$ for the data pair (x, y).
$f(x; w)$ is the *regression function* and the vector w are its parameters.

Definition of linear regression

A regression method is called *linear* if the prediction f is a linear function of the unknown parameters w.

LEAST SQUARES (CONTINUED)

Least squares solution

Least squares finds the w that minimizes the sum of squared errors. The least squares objective in the most basic form where $f(x; w) = x^T w$ is

$$\mathcal{L} = \sum_{i=1}^{n} (y_i - x_i^T w)^2 \;=\; \|y - Xw\|^2 \;=\; (y - Xw)^T (y - Xw).$$

We defined $y = [y_1, \ldots, y_n]^T$ and $X = [x_1, \ldots, x_n]^T$.

Taking the gradient with respect to w and setting to zero, we find that

$$\nabla_w \mathcal{L} = 2X^T Xw - 2X^T y = 0 \quad \Rightarrow \quad w_{\text{LS}} = (X^T X)^{-1} X^T y.$$

In other words, w_{LS} is the vector that minimizes $\mathcal{L}$.

PROBABILISTIC VIEW

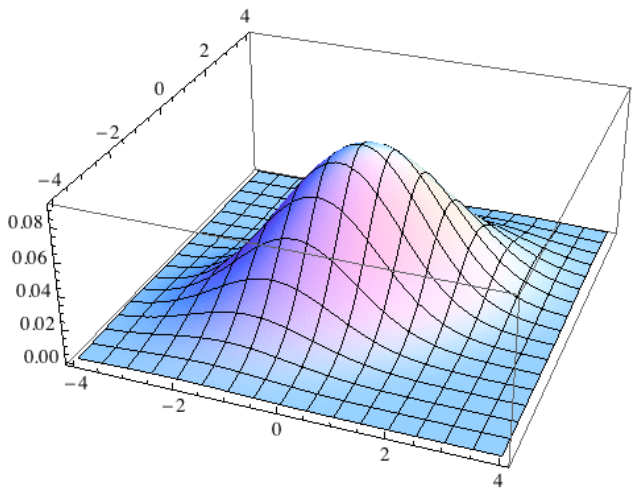
- Last class, we discussed the geometric interpretation of least squares.
- Least squares also has an insightful probabilistic interpretation that allows us to analyze its properties.
- That is, given that we pick this model as reasonable for our problem, we can ask: What kinds of assumptions are we making?

Recall: Gaussian density in n dimensions

Assume a diagonal covariance matrix $\Sigma = \sigma^2 I$. The density is

$$p(y|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\Big(-\frac{1}{2\sigma^2}(y-\mu)^T(y-\mu)\Big).$$

What if we restrict the mean to $\mu = Xw$ and find the *maximum likelihood* solution for w?



Maximum likelihood for Gaussian linear regression

Plug $\mu = Xw$ into the multivariate Gaussian distribution and solve for w using maximum likelihood.

$$\begin{aligned} w_{\text{ML}} &= \arg\max_w \ln p(y|\mu = Xw, \sigma^2) \\ &= \arg\max_w -\frac{1}{2\sigma^2}\|y - Xw\|^2 - \frac{n}{2}\ln(2\pi\sigma^2). \end{aligned}$$

Least squares (LS) and maximum likelihood (ML) share the same solution:

$$\text{LS:} \quad \arg\min_w \|y - Xw\|^2 \quad \Leftrightarrow \quad \text{ML:} \quad \arg\max_w -\frac{1}{2\sigma^2}\|y - Xw\|^2$$

PROBABILISTIC VIEW

- Therefore, in a sense we are making an *independent Gaussian noise* assumption about the error, $\epsilon_i = y_i - x_i^T w$.

- Other ways of saying this:

  1) $y_i = x_i^T w + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \quad \text{for } i = 1, \ldots, n,$

  2) $y_i \overset{ind}{\sim} N(x_i^T w, \sigma^2), \quad \text{for } i = 1, \ldots, n,$

  3) $y \sim N(Xw, \sigma^2 I),$ as on the previous slides.

- Can we use this probabilistic line of analysis to better understand the maximum likelihood (i.e., least squares) solution?

Expected solution

Given: The *modeling assumption* that $y \sim N(Xw, \sigma^2 I)$.

We can calculate the expectation of the ML solution under this distribution,

$$\begin{aligned}
\mathbb{E}[w_{\text{ML}}] &= \mathbb{E}[(X^T X)^{-1} X^T y] \qquad \left(= \int \left[(X^T X)^{-1} X^T y\right] p(y|X, w)\, dy \right) \\
&= (X^T X)^{-1} X^T \mathbb{E}[y] \\
&= (X^T X)^{-1} X^T X w \\
&= w
\end{aligned}$$

Therefore w_{ML} is an *unbiased* estimate of w, i.e., $\mathbb{E}[w_{\text{ML}}] = w$.

Review: An equality from probability

- Even though the "expected" maximum likelihood solution is the correct one, should we actually expect to get something near it?

Review: An equality from probability

- Even though the "expected" maximum likelihood solution is the correct one, should we actually expect to get something near it?

- We should also look at the covariance. Recall that if $y \sim N(\mu, \Sigma)$, then

$$\text{Var}[y] = \mathbb{E}[(y - \mathbb{E}[y])(y - \mathbb{E}[y])^T] = \Sigma.$$

REVIEW: AN EQUALITY FROM PROBABILITY

- Even though the "expected" maximum likelihood solution is the correct one, should we actually expect to get something near it?

- We should also look at the covariance. Recall that if $y \sim N(\mu, \Sigma)$, then

$$\text{Var}[y] = \mathbb{E}[(y - \mathbb{E}[y])(y - \mathbb{E}[y])^T] = \Sigma.$$

- Plugging in $\mathbb{E}[y] = \mu$, this is equivalently written as

$$\begin{aligned} \text{Var}[y] &= \mathbb{E}[(y - \mu)(y - \mu)^T] \\ &= \mathbb{E}[yy^T - y\mu^T - \mu y^T + \mu\mu^T] \\ &= \mathbb{E}[yy^T] - \mu\mu^T \end{aligned}$$

- Immediately we also get $\mathbb{E}[yy^T] = \Sigma + \mu\mu^T$.

Variance of the solution

Returning to least squares linear regression, we wish to find

$$\begin{aligned} \mathrm{Var}[w_{\mathrm{ML}}] &= \mathbb{E}[(w_{\mathrm{ML}} - \mathbb{E}[w_{\mathrm{ML}}])(w_{\mathrm{ML}} - \mathbb{E}[w_{\mathrm{ML}}])^T] \\ &= \mathbb{E}[w_{\mathrm{ML}} w_{\mathrm{ML}}^T] - \mathbb{E}[w_{\mathrm{ML}}]\mathbb{E}[w_{\mathrm{ML}}]^T. \end{aligned}$$

[1] Aside: For matrices A, B and vector c, recall that $(ABc)^T = c^T B^T A^T$.

Variance of the solution

Returning to least squares linear regression, we wish to find

$$\begin{aligned} \mathrm{Var}[w_{\mathrm{ML}}] &= \mathbb{E}[(w_{\mathrm{ML}} - \mathbb{E}[w_{\mathrm{ML}}])(w_{\mathrm{ML}} - \mathbb{E}[w_{\mathrm{ML}}])^T] \\ &= \mathbb{E}[w_{\mathrm{ML}} w_{\mathrm{ML}}^T] - \mathbb{E}[w_{\mathrm{ML}}]\mathbb{E}[w_{\mathrm{ML}}]^T. \end{aligned}$$

The sequence of equalities follows:[1]

$$\mathrm{Var}[w_{\mathrm{ML}}] = \mathbb{E}[(X^T X)^{-1} X^T y y^T X (X^T X)^{-1}] - w w^T$$

[1] Aside: For matrices A, B and vector c, recall that $(ABc)^T = c^T B^T A^T$.

Variance of the solution

Returning to least squares linear regression, we wish to find

$$\begin{aligned} \mathrm{Var}[w_{\mathrm{ML}}] &= \mathbb{E}[(w_{\mathrm{ML}} - \mathbb{E}[w_{\mathrm{ML}}])(w_{\mathrm{ML}} - \mathbb{E}[w_{\mathrm{ML}}])^T] \\ &= \mathbb{E}[w_{\mathrm{ML}} w_{\mathrm{ML}}^T] - \mathbb{E}[w_{\mathrm{ML}}]\mathbb{E}[w_{\mathrm{ML}}]^T. \end{aligned}$$

The sequence of equalities follows:[1]

$$\begin{aligned} \mathrm{Var}[w_{\mathrm{ML}}] &= \mathbb{E}[(X^T X)^{-1} X^T y y^T X (X^T X)^{-1}] - w w^T \\ &= (X^T X)^{-1} X^T \mathbb{E}[y y^T] X (X^T X)^{-1} - w w^T \end{aligned}$$

[1] Aside: For matrices A, B and vector c, recall that $(ABc)^T = c^T B^T A^T$.

Variance of the solution

Returning to least squares linear regression, we wish to find

$$\begin{aligned}\mathrm{Var}[w_{\mathrm{ML}}] &= \mathbb{E}[(w_{\mathrm{ML}} - \mathbb{E}[w_{\mathrm{ML}}])(w_{\mathrm{ML}} - \mathbb{E}[w_{\mathrm{ML}}])^T] \\ &= \mathbb{E}[w_{\mathrm{ML}} w_{\mathrm{ML}}^T] - \mathbb{E}[w_{\mathrm{ML}}]\mathbb{E}[w_{\mathrm{ML}}]^T.\end{aligned}$$

The sequence of equalities follows:[1]

$$\begin{aligned}\mathrm{Var}[w_{\mathrm{ML}}] &= \mathbb{E}[(X^TX)^{-1}X^T yy^T X(X^TX)^{-1}] - ww^T \\ &= (X^TX)^{-1}X^T\mathbb{E}[yy^T]X(X^TX)^{-1} - ww^T \\ &= (X^TX)^{-1}X^T(\sigma^2 I + Xww^TX^T)X(X^TX)^{-1} - ww^T\end{aligned}$$

[1] Aside: For matrices A, B and vector c, recall that $(ABc)^T = c^TB^TA^T$.

Variance of the solution

Returning to least squares linear regression, we wish to find

$$\begin{aligned}\mathrm{Var}[w_{\mathrm{ML}}] &= \mathbb{E}[(w_{\mathrm{ML}} - \mathbb{E}[w_{\mathrm{ML}}])(w_{\mathrm{ML}} - \mathbb{E}[w_{\mathrm{ML}}])^T] \\ &= \mathbb{E}[w_{\mathrm{ML}} w_{\mathrm{ML}}^T] - \mathbb{E}[w_{\mathrm{ML}}]\mathbb{E}[w_{\mathrm{ML}}]^T.\end{aligned}$$

The sequence of equalities follows:[1]

$$\begin{aligned}\mathrm{Var}[w_{\mathrm{ML}}] &= \mathbb{E}[(X^TX)^{-1}X^Tyy^TX(X^TX)^{-1}] - ww^T \\ &= (X^TX)^{-1}X^T\mathbb{E}[yy^T]X(X^TX)^{-1} - ww^T \\ &= (X^TX)^{-1}X^T(\sigma^2 I + Xww^TX^T)X(X^TX)^{-1} - ww^T \\ &= (X^TX)^{-1}X^T\sigma^2 IX(X^TX)^{-1} + \cdots \\ &\quad (X^TX)^{-1}X^TXww^TX^TX(X^TX)^{-1} - ww^T\end{aligned}$$

[1] Aside: For matrices A, B and vector c, recall that $(ABc)^T = c^TB^TA^T$.

Variance of the solution

Returning to least squares linear regression, we wish to find

$$\begin{aligned}\mathrm{Var}[w_{\mathrm{ML}}] &= \mathbb{E}[(w_{\mathrm{ML}} - \mathbb{E}[w_{\mathrm{ML}}])(w_{\mathrm{ML}} - \mathbb{E}[w_{\mathrm{ML}}])^T] \\ &= \mathbb{E}[w_{\mathrm{ML}} w_{\mathrm{ML}}^T] - \mathbb{E}[w_{\mathrm{ML}}]\mathbb{E}[w_{\mathrm{ML}}]^T.\end{aligned}$$

The sequence of equalities follows:[1]

$$\begin{aligned}\mathrm{Var}[w_{\mathrm{ML}}] &= \mathbb{E}[(X^TX)^{-1}X^Tyy^TX(X^TX)^{-1}] - ww^T \\ &= (X^TX)^{-1}X^T\mathbb{E}[yy^T]X(X^TX)^{-1} - ww^T \\ &= (X^TX)^{-1}X^T(\sigma^2 I + Xww^TX^T)X(X^TX)^{-1} - ww^T \\ &= (X^TX)^{-1}X^T\sigma^2 IX(X^TX)^{-1} + \cdots \\ &\quad (X^TX)^{-1}X^TXww^TX^TX(X^TX)^{-1} - ww^T \\ &= \sigma^2(X^TX)^{-1}\end{aligned}$$

[1] Aside: For matrices A, B and vector c, recall that $(ABc)^T = c^TB^TA^T$.

PROBABILISTIC VIEW

- We've shown that, under the Gaussian assumption $y \sim N(Xw, \sigma^2 I)$,

$$\mathbb{E}[w_{\text{ML}}] = w, \quad \text{Var}[w_{\text{ML}}] = \sigma^2 (X^T X)^{-1}.$$

- When there are very large values in $\sigma^2 (X^T X)^{-1}$, the values of w_{ML} are very sensitive to the measured data y (more analysis later).

- This is bad if we want to analyze and predict using w_{ML}.

Ridge Regression

REGULARIZED LEAST SQUARES

- We saw how with least squares, the values in w_{ML} may be huge.
- In general, when developing a model for data we often wish to *constrain* the model parameters in some way.
- There are many models of the form

$$w_{\text{OPT}} = \arg\min_{w} \|y - Xw\|^2 + \lambda\, g(w).$$

- The added terms are
  1. $\lambda > 0$: a regularization parameter,
  2. $g(w) > 0$: a penalty function that encourages desired properties about w.

RIDGE REGRESSION

Ridge regression is one $g(w)$ that addresses variance issues with w_{ML}.

It uses the squared penalty on the regression coefficient vector w,

$$w_{\text{RR}} \quad = \quad \arg\min_{w} \ \|y - Xw\|^2 + \lambda \|w\|^2$$

The term $g(w) = \|w\|^2$ penalizes large values in w.

However, there is a *tradeoff* between the first and second terms that is controlled by λ.

- Case $\lambda \to 0 \quad : \ w_{\text{RR}} \to w_{\text{LS}}$
- Case $\lambda \to \infty \ : \ w_{\text{RR}} \to \vec{0}$

RIDGE REGRESSION SOLUTION

Objective: We can solve the ridge regression problem using exactly the same procedure as for least squares,

$$\begin{aligned}\mathcal{L} &= \|y - Xw\|^2 + \lambda\|w\|^2 \\ &= (y - Xw)^T(y - Xw) + \lambda w^T w.\end{aligned}$$

Solution: First, take the gradient of $\mathcal{L}$ with respect to w and set to zero,

$$\nabla_w \mathcal{L} = -2X^T y + 2X^T Xw + 2\lambda w = 0$$

Then, solve for w to find that

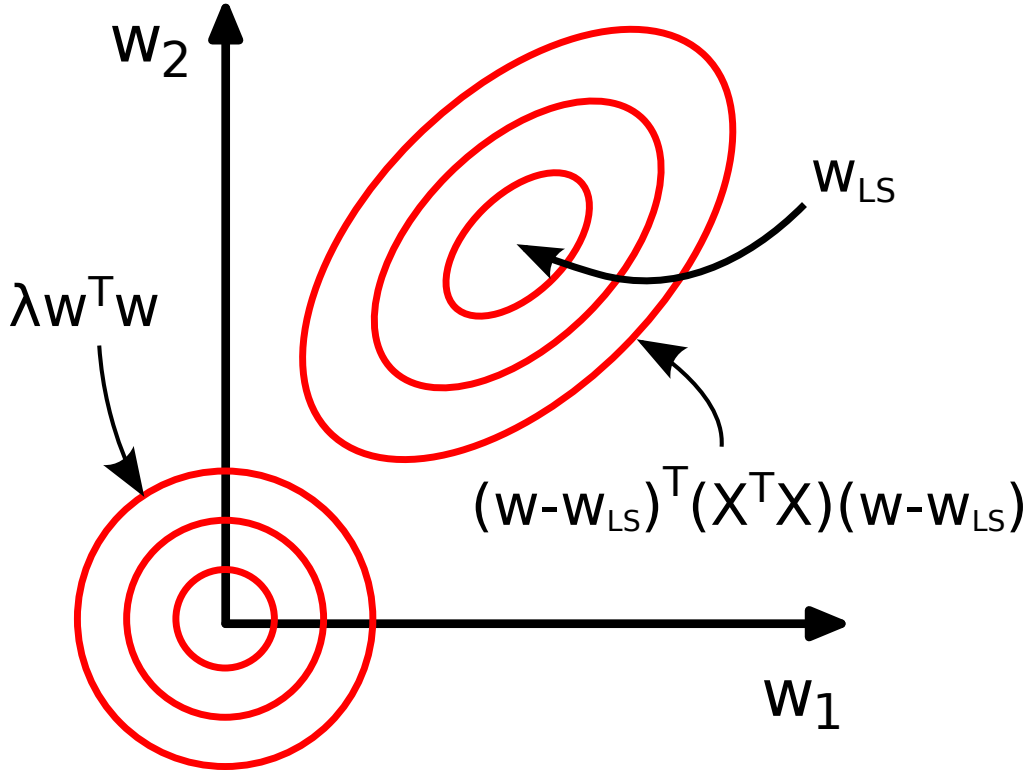
$$w_{\text{RR}} = (\lambda I + X^T X)^{-1} X^T y.$$

RIDGE REGRESSION GEOMETRY

There is a tradeoff between squared error and penalty on w.

We can write both in terms of *level sets*: Curves where function evaluation gives the same number.

The sum of these gives a new set of levels with a unique minimum.



You can check that we can write:

$$\|y - Xw\|^2 + \lambda\|w\|^2 = (w - w_{\text{LS}})^T(X^TX)(w - w_{\text{LS}}) + \lambda w^T w + (\text{const. w.r.t. } w).$$

Data preprocessing

Ridge regression is one possible regularization scheme. For this problem, we first assume the following *preprocessing* steps are done:

1. The mean is subtracted off of y:

$$y \leftarrow y - \frac{1}{n}\sum_{i=1}^{n} y_i.$$

2. The dimensions of x_i have been *standardized* before constructing X:

$$x_{ij} \leftarrow (x_{ij} - \bar{x}_{\cdot j})/\hat{\sigma}_j, \quad \hat{\sigma}_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_{\cdot j})^2}.$$

   i.e., subtract the empirical mean and divide by the empirical standard deviation for each dimension.

3. We can show that there is no need for the dimension of 1's in this case.

Some Analysis of Ridge Regression

Ridge regression vs Least squares

The solutions to least squares and ridge regression are clearly very similar,

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

- We can use linear algebra and probability to compare the two.
- This requires the *singular value decomposition*, which we review next.

REVIEW: SINGULAR VALUE DECOMPOSITIONS

- We can write any $n \times d$ matrix X (assume $n > d$) as $X = USV^T$, where
  1. U: $n \times d$ and orthonormal in the columns, i.e. $U^T U = I$.
  2. S: $d \times d$ non-negative diagonal matrix, i.e. $S_{ii} \geq 0$ and $S_{ij} = 0$ for $i \neq j$.
  3. V: $d \times d$ and orthonormal, i.e. $V^T V = VV^T = I$.

- From this we have the immediate equalities

$$X^T X = (USV^T)^T (USV^T) = VS^2 V^T, \quad XX^T = US^2 U^T.$$

- Assuming $S_{ii} \neq 0$ for all i (i.e., "X is full rank"), we also have that

$$(X^T X)^{-1} = (VS^2 V^T)^{-1} = VS^{-2} V^T.$$

  Proof: Plug in and see that it satisfies definition of inverse

$$(X^T X)(X^T X)^{-1} = VS^2 V^T VS^{-2} V^T = I.$$

Least squares and the SVD

Using the SVD we can rewrite the variance,

$$\mathrm{Var}[w_{\mathrm{LS}}] = \sigma^2 (X^T X)^{-1} = \sigma^2 V S^{-2} V^T.$$

This inverse becomes huge when S_{ii} is very small for some values of i.
(Aside: This happens when columns of X are highly correlated.)

The least squares prediction for new data is

$$y_{\mathrm{new}} = x_{\mathrm{new}}^T w_{\mathrm{LS}} = x_{\mathrm{new}}^T (X^T X)^{-1} X^T y = x_{\mathrm{new}}^T V S^{-1} U^T y.$$

When S^{-1} has very large values, this can lead to unstable predictions.

Relationship to least squares solution

Recall for two symmetric matrices, $(AB)^{-1} = B^{-1}A^{-1}$.

$$
\begin{aligned}
w_{\text{RR}} &= (\lambda I + X^T X)^{-1} X^T y \\
&= (\lambda I + X^T X)^{-1} (X^T X) \underbrace{(X^T X)^{-1} X^T y}_{w_{\text{LS}}} \\
&= [(X^T X)(\lambda (X^T X)^{-1} + I)]^{-1} (X^T X) w_{\text{LS}} \\
&= (\lambda (X^T X)^{-1} + I)^{-1} (X^T X)^{-1} (X^T X) w_{\text{LS}} \\
&= (\lambda (X^T X)^{-1} + I)^{-1} w_{\text{LS}}
\end{aligned}
$$

Can use this to prove that the solution shrinks toward zero: $\|w_{\text{RR}}\|_2 \leq \|w_{\text{LS}}\|_2$.

RIDGE REGRESSION VS LEAST SQUARES II

Continue analysis with the SVD: $X = USV^T \rightarrow (X^TX)^{-1} = VS^{-2}V^T$:

$$\begin{aligned} w_{\text{RR}} &= (\lambda(X^TX)^{-1} + I)^{-1} w_{\text{LS}} \\ &= (\lambda VS^{-2}V^T + I)^{-1} w_{\text{LS}} \\ &= V(\lambda S^{-2} + I)^{-1} V^T w_{\text{LS}} \\ &:= VMV^T w_{\text{LS}} \end{aligned}$$

M is a diagonal matrix with $M_{ii} = \frac{S_{ii}^2}{\lambda + S_{ii}^2}$. We can pursue this to show that

$$w_{\text{RR}} = VS_\lambda^{-1}U^T y, \quad S_\lambda^{-1} = \begin{bmatrix} \frac{S_{11}}{\lambda + S_{11}^2} & & 0 \\ & \ddots & \\ 0 & & \frac{S_{dd}}{\lambda + S_{dd}^2} \end{bmatrix}$$

Compare with $w_{\text{LS}} = VS^{-1}U^T y$, which is the case where $\lambda = 0$ above.

RIDGE REGRESSION VS LEAST SQUARES III

Ridge regression can also be seen as a special case of least squares.

Define $\hat{y} \approx \hat{X}w$ in the following way,

$$\begin{bmatrix} y \\ 0 \\ \vdots \\ 0 \end{bmatrix} \approx \begin{bmatrix} - & X & - \\ \sqrt{\lambda} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda} \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}$$

If we solved w_{LS} for *this* regression problem, we find w_{RR} of the *original* problem: Calculating $(\hat{y} - \hat{X}w)^T(\hat{y} - \hat{X}w)$ in two parts gives

$$\begin{aligned} (\hat{y} - \hat{X}w)^T(\hat{y} - \hat{X}w) &= (y - Xw)^T(y - Xw) + (\sqrt{\lambda}w)^T(\sqrt{\lambda}w) \\ &= \|y - Xw\|^2 + \lambda\|w\|^2 \end{aligned}$$

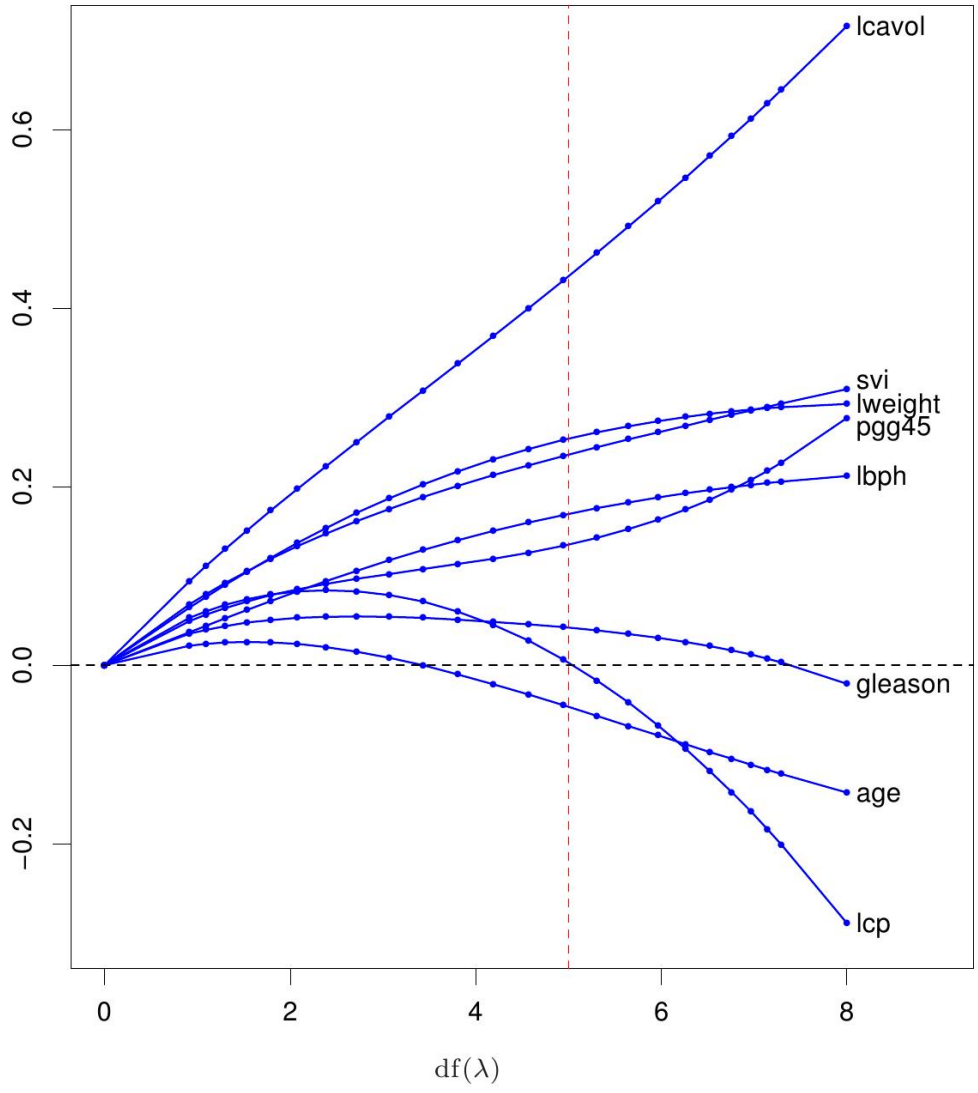
SELECTING λ

Degrees of freedom:

$$df(\lambda) = \text{trace}\left[X(X^TX + \lambda I)^{-1}X^T\right]$$
$$= \sum_{i=1}^{d} \frac{S_{ii}^2}{\lambda + S_{ii}^2}$$

This gives a way of visualizing relationships.

We will discuss methods for picking λ later.



ColumbiaX: Machine Learning

Lecture 4

Prof. John Paisley

Department of Electrical Engineering
& Data Science Institute

Columbia University

REGRESSION WITH/WITHOUT REGULARIZATION

Given:

A data set $(x_1, y_1), \ldots, (x_n, y_n)$, where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$. We standardize such that each dimension of x is zero mean unit variance, and y is zero mean.

Model:

We define a model of the form

$$y \approx f(x; w).$$

We particularly focus on the case where $f(x; w) = x^T w$.

Learning:

We can learn the model by minimizing the objective (aka, "loss") function

$$\mathcal{L} = \sum_{i=1}^{n} (y_i - x_i^T w)^2 + \lambda w^T w \quad \Leftrightarrow \quad \mathcal{L} = \|y - Xw\|^2 + \lambda \|w\|^2$$

We've focused on $\lambda = 0$ (least squares) and $\lambda > 0$ (ridge regression).

BIAS-VARIANCE TRADE-OFF

BIAS-VARIANCE FOR LINEAR REGRESSION

We can go further and hypothesize a *generative* model $y \sim N(Xw, \sigma^2 I)$ and some true (but unknown) underlying value for the parameter vector w.

- We saw how the least squares solution, $w_{\text{LS}} = (X^T X)^{-1} X^T y$, is unbiased but potentially has high variance:

$$\mathbb{E}[w_{\text{LS}}] = w, \quad \text{Var}[w_{\text{LS}}] = \sigma^2 (X^T X)^{-1}.$$

- By contrast, the ridge regression solution is $w_{\text{RR}} = (\lambda I + X^T X)^{-1} X^T y$. Using the same procedure as for least squares, we can show that

$$\mathbb{E}[w_{\text{RR}}] = (\lambda I + X^T X)^{-1} X^T X w, \quad \text{Var}[w_{\text{RR}}] = \sigma^2 Z (X^T X)^{-1} Z^T,$$

  where $Z = (I + \lambda (X^T X)^{-1})^{-1}$.

BIAS-VARIANCE FOR LINEAR REGRESSION

The expectation and covariance of w_{LS} and w_{RR} gives insight into how well we can hope to learn w in the case where our model assumption is correct.

- Least squares solution: unbiased, but potentially high variance
- Ridge regression solution: biased, but lower variance than LS

So which is preferable?

Ultimately, we really care about how well our solution for w generalizes to new data. Let (x_0, y_0) be future data for which we have x_0, but not y_0.

- Least squares predicts $y_0 = x_0^T w_{\text{LS}}$
- Ridge regression predicts $y_0 = x_0^T w_{\text{RR}}$

BIAS-VARIANCE FOR LINEAR REGRESSION

In keeping with the square error measure of performance, we could calculate the expected squared error of our prediction:

$$\mathbb{E}\left[(y_0 - x_0^T\hat{w})^2 | X, x_0\right] = \int_{\mathbb{R}} \int_{\mathbb{R}^n} (y_0 - x_0^T\hat{w})^2 p(y|X,w) p(y_0|x_0,w)\, dy\, dy_0.$$

- The estimate $\hat{w}$ is either w_{LS} or w_{RR}.
- The distributions on y, y_0 are Gaussian with the true (but unknown) w.
- We condition on knowing $x_0, x_1, \ldots, x_n$.

In words this is saying:

- Imagine I know X, x_0 and assume some true underlying w.
- I generate $y \sim N(Xw, \sigma^2 I)$ and approximate w with $\hat{w} = w_{\text{LS}}$ or w_{RR}.
- I then predict $y_0 \sim N(x_0^T w, \sigma^2)$ using $y_0 \approx x_0^T \hat{w}$.

What is the expected squared error of my prediction?

BIAS-VARIANCE FOR LINEAR REGRESSION

We can calculate this as follows (assume conditioning on x_0 and X),

$$\mathbb{E}[(y_0 - x_0^T \hat{w})^2] = \mathbb{E}[y_0^2] - 2\mathbb{E}[y_0] x_0^T \mathbb{E}[\hat{w}] + x_0^T \mathbb{E}[\hat{w}\hat{w}^T] x_0$$

- Since y_0 and $\hat{w}$ are independent, $\mathbb{E}[y_0 \hat{w}] = \mathbb{E}[y_0]\mathbb{E}[\hat{w}]$.
- Remember: $\mathbb{E}[\hat{w}\hat{w}^T] = \text{Var}[\hat{w}] + \mathbb{E}[\hat{w}]\mathbb{E}[\hat{w}]^T$

$$\mathbb{E}[y_0^2] = \sigma^2 + (x_0^T w)^2$$

BIAS-VARIANCE FOR LINEAR REGRESSION

We can calculate this as follows (assume conditioning on x_0 and X),

$$\mathbb{E}[(y_0 - x_0^T \hat{w})^2] = \mathbb{E}[y_0^2] - 2\mathbb{E}[y_0] x_0^T \mathbb{E}[\hat{w}] + x_0^T \mathbb{E}[\hat{w}\hat{w}^T] x_0$$

- Since y_0 and $\hat{w}$ are independent, $\mathbb{E}[y_0 \hat{w}] = \mathbb{E}[y_0]\mathbb{E}[\hat{w}]$.
- Remember: $\mathbb{E}[\hat{w}\hat{w}^T] = \text{Var}[\hat{w}] + \mathbb{E}[\hat{w}]\mathbb{E}[\hat{w}]^T$

$$\mathbb{E}[y_0^2] = \sigma^2 + (x_0^T w)^2$$

Plugging these values in:

$$\begin{aligned}
\mathbb{E}[(y_0 - x_0^T \hat{w})^2] &= \sigma^2 + (x_0^T w)^2 - 2(x_0^T w)(x_0^T \mathbb{E}[\hat{w}]) + (x_0^T \mathbb{E}[\hat{w}])^2 + x_0^T \text{Var}[\hat{w}] x_0 \\
&= \sigma^2 + x_0^T (w - \mathbb{E}[\hat{w}])(w - \mathbb{E}[\hat{w}])^T x_0 + x_0^T \text{Var}[\hat{w}] x_0
\end{aligned}$$

BIAS-VARIANCE FOR LINEAR REGRESSION

We have shown that if

1. $y \sim N(Xw, \sigma^2)$ and $y_0 \sim N(x_0^T w, \sigma^2)$, and
2. we approximate w with $\hat{w}$ according to some algorithm,

then

$$\mathbb{E}[(y_0 - x_0^T \hat{w})^2 | X, x_0] = \underbrace{\sigma^2}_{\textit{noise}} + \underbrace{x_0^T (w - \mathbb{E}[\hat{w}])(w - \mathbb{E}[\hat{w}])^T x_0}_{\textit{squared bias}} + \underbrace{x_0^T \mathrm{Var}[\hat{w}] x_0}_{\textit{variance}}$$

We see that the *generalization error* is a combination of three factors:

1. Measurement noise – we can't control this given the model.
2. Model bias – how close to the solution we expect to be on average.
3. Model variance – how sensitive our solution is to the data.

We saw how we can find $\mathbb{E}[\hat{w}]$ and $\mathrm{Var}[\hat{w}]$ for the LS and RR solutions.

BIAS-VARIANCE TRADE-OFF

This idea is more general:

- Imagine we have a model: $y = f(x; w) + \epsilon,\ \mathbb{E}(\epsilon) = 0,\ \mathrm{Var}(\epsilon) = \sigma^2$
- We approximate f by minimizing a loss function: $\hat{f} = \arg\min_f \mathcal{L}_f$.
- We apply $\hat{f}$ to new data, $y_0 \approx \hat{f}(x_0) \equiv \hat{f}_0$.

Then integrating everything out (y, X, y_0, x_0):

$$\begin{aligned}
\mathbb{E}[(y_0 - \hat{f}_0)^2] &= \mathbb{E}[y_0^2] - 2\mathbb{E}[y_0 \hat{f}_0] + \mathbb{E}[\hat{f}_0^2] \\
&= \sigma^2 + f_0^2 - 2f_0\mathbb{E}[\hat{f}_0] + \mathbb{E}[\hat{f}_0]^2 + \mathrm{Var}[\hat{f}_0] \\
&= \underbrace{\sigma^2}_{\textit{noise}} + \underbrace{(f_0 - \mathbb{E}[\hat{f}_0])^2}_{\textit{squared bias}} + \underbrace{\mathrm{Var}[\hat{f}_0]}_{\textit{variance}}
\end{aligned}$$

This is interesting in principle, but is deliberately vague (What is f?) and usually can't be calculated (What is the distribution on the data?)

CROSS-VALIDATION

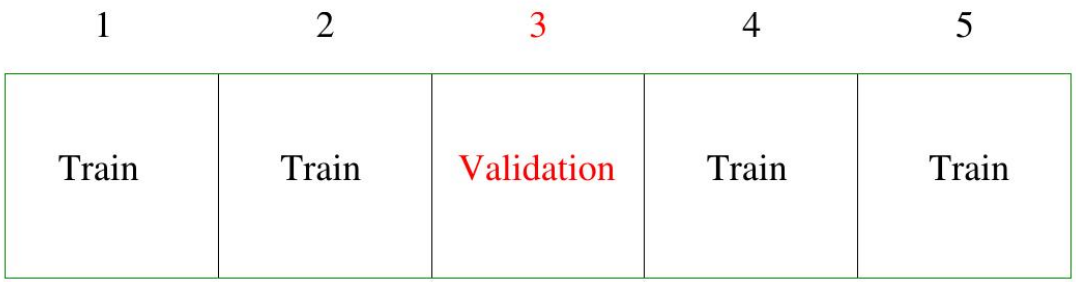
An easier way to evaluate the model is to use cross-validation.

The procedure for K-fold cross-validation is very simple:

1. Randomly split the data into K roughly equal groups.
2. Learn the model on $K - 1$ groups and predict the held-out Kth group.
3. Do this K times, holding out each group once.
4. Evaluate performance using the cumulative set of predictions.

For the case of the regularization parameter λ, the above sequence can be run for several values with the best-performing value of λ chosen.

The data you test the model on should never be used to train the model!



BAYES RULE

Motivation

We've discussed the ridge regression objective function

$$\mathcal{L} = \sum_{i=1}^{n} (y_i - x_i^T w)^2 + \lambda w^T w.$$

The regularization term $\lambda w^T w$ was imposed to penalize values in w that are large. This reduced potential high-variance predictions from least squares.

In a sense, we are imposing a "prior belief" about what values of w we consider to be good.

Question: Is there a mathematical way to formalize this?

Answer: Using probability we can frame this via Bayes rule.

Review: Probability statements

Imagine we have two events, A and B, that may or may not be related, e.g.,

- A = "It is raining"
- B = "The ground is wet"

We can talk about probabilities of these events,

- $P(A)$ = Probability it is raining
- $P(B)$ = Probability the ground is wet

We can also talk about their *conditional* probabilities,

- $P(A|B)$ = Probability it is raining *given* that the ground is wet
- $P(B|A)$ = Probability the ground is wet *given* that it is raining

We can also talk about their *joint* probabilities,

- $P(A, B)$ = Probability it is raining *and* the ground is wet

CALCULUS OF PROBABILITY

There are simple rules from moving from one probability to another

1. $P(A,B) = P(A|B)P(B) = P(B|A)P(A)$
2. $P(A) = \sum_b P(A, B = b)$
3. $P(B) = \sum_a P(A = a, B)$

Using these three equalities, we automatically can say

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{\sum_a P(B|A = a)P(A = a)}$$

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{P(A|B)P(B)}{\sum_b P(A|B = b)P(B = b)}$$

This is known as “Bayes rule.”

BAYES RULE

Bayes rule lets us quantify what we don't know. Imagine we want to say something about the probability of B given that A happened.

Bayes rule says that the probability of B after knowing A is:

$$\underbrace{P(B|A)}_{posterior} = \underbrace{P(A|B)}_{likelihood}\underbrace{P(B)}_{prior} \; / \; \underbrace{P(A)}_{marginal}$$

Notice that with this perspective, these probabilities take on new meanings.

That is, $P(B|A)$ and $P(A|B)$ are both "conditional probabilities," but they have different significance.

BAYES RULE WITH CONTINUOUS VARIABLES

Bayes rule generalizes to continuous-valued random variables as follows. However, instead of *probabilities* we work with *densities*.

- Let θ be a continuous-valued model parameter.
- Let X be data we possess. Then by Bayes rule,

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{\int p(X|\theta)p(\theta)d\theta} = \frac{p(X|\theta)p(\theta)}{p(X)}$$

In this equation,

- $p(X|\theta)$ is the likelihood, known from the model definition.
- $p(\theta)$ is a prior distribution that we define.
- Given these two, we can (in principle) calculate $p(\theta|X)$.

EXAMPLE: COIN BIAS

We have a coin with bias π towards "heads". (Encode: heads = 1, tails = 0)

We flip the coin many times and get a sequence of n numbers $(x_1, \ldots, x_n)$. Assume the flips are independent, meaning

$$p(x_1, \ldots, x_n|\pi) = \prod_{i=1}^{n} p(x_i|\pi) = \prod_{i=1}^{n} \pi^{x_i}(1-\pi)^{1-x_i}.$$

We choose a prior for π which we define to be a beta distribution,

$$p(\pi) = \mathit{Beta}(\pi|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\pi^{a-1}(1-\pi)^{b-1}.$$

What is the posterior distribution of π given $x_1, \ldots, x_n$?

EXAMPLE: COIN BIAS

From Bayes rule,

$$p(\pi|x_1,\ldots,x_n) = \frac{p(x_1,\ldots,x_n|\pi)p(\pi)}{\int_0^1 p(x_1,\ldots,x_n|\pi)p(\pi)d\pi}.$$

There is a trick that is often useful:

- The denominator only normalizes the numerator, doesn't depend on π.
- We can write $p(\pi|x) \propto p(x|\pi)p(\pi)$. ("$\propto$" $\rightarrow$ "proportional to")
- Multiply the two and see if we recognize anything:

$$\begin{aligned} p(\pi|x_1,\ldots,x_n) &\propto \left[\prod_{i=1}^n \pi^{x_i}(1-\pi)^{1-x_i}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\pi^{a-1}(1-\pi)^{b-1}\right] \\ &\propto \pi^{\sum_{i=1}^n x_i + a - 1}(1-\pi)^{\sum_{i=1}^n (1-x_i)+b-1} \end{aligned}$$

We recognize this as $p(\pi|x_1,\ldots,x_n) = Beta(\sum_{i=1}^n x_i + a, \sum_{i=1}^n (1-x_i) + b)$.

Maximum a posteriori

Least squares and maximum likelihood

When we modeled data pairs (x_i, y_i) with a linear model, $y_i \approx x_i^T w$, we saw that the least squares solution,

$$w_{\text{LS}} = \arg\min_w (y - Xw)^T (y - Xw),$$

was equivalent to the maximum likelihood solution when $y \sim N(Xw, \sigma^2 I)$.

The question now is whether a similar probabilistic connection can be made for the ridge regression problem.

Ridge regression and Bayesian modeling

The likelihood model is $y \sim N(Xw, \sigma^2 I)$. What about a prior for w?

Let us assume that the prior for w is Gaussian, $w \sim N(0, \lambda^{-1} I)$. Then

$$p(w) = \left(\frac{\lambda}{2\pi}\right)^{\frac{d}{2}} \mathrm{e}^{-\frac{\lambda}{2} w^T w}.$$

We can now try to find a w that satisfies both the data likelihood, and our prior conditions about w.

Maximum a poseriori estimation

Maximum *a poseriori* (MAP) estimation seeks the most probable value w under the posterior:

$$\begin{aligned} w_{\text{MAP}} &= \arg\max_w \ln p(w|y,X) \\ &= \arg\max_w \ln \frac{p(y|w,X)p(w)}{p(y|X)} \\ &= \arg\max_w \ln p(y|w,X) + \ln p(w) - \ln p(y|X) \end{aligned}$$

- Contrast this with ML, which only focuses on the likelihood.
- The normalizing constant term $\ln p(y|X)$ doesn't involve w. Therefore, we can maximize the first two terms alone.
- In many models we don't know $\ln p(y|X)$, so this fact is useful.

MAP FOR LINEAR REGRESSION

MAP using our defined prior gives:

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

Calling this objective $\mathcal{L}$, then as before we find w such that

$$\nabla_w \mathcal{L} = \frac{1}{\sigma^2}X^T y - \frac{1}{\sigma^2}X^T Xw - \lambda w = 0$$

- The solution is $w_{\text{MAP}} = (\lambda\sigma^2 I + X^T X)^{-1}X^T y$.
- Notice that $w_{\text{MAP}} = w_{\text{RR}}$ (modulo a switch from λ to $\lambda\sigma^2$)
- RR maximizes the posterior, while LS maximizes the likelihood.