COMS 4721: Machine Learning for Data Science
Lecture 3, 1/24/2017

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

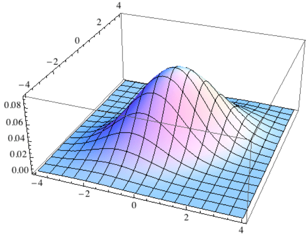
- Last class, we discussed the geometric interpretation of least squares.
- Least squares also has an insightful probabilistic interpretation that allows us to analyze its properties.
- That is, given that we pick this model as reasonable for our problem, we can ask: What kinds of assumptions are we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

Recall: Gaussian density in n dimensions

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

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

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

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

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

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

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

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

- 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]

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

[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]

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

[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

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

The sequence of equalities follows:[1]

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

[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,

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

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 $\lambda \to 0$: $w_{\text{RR}} \to w_{\text{LS}}$
- 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,

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

Then, solve for w to find that

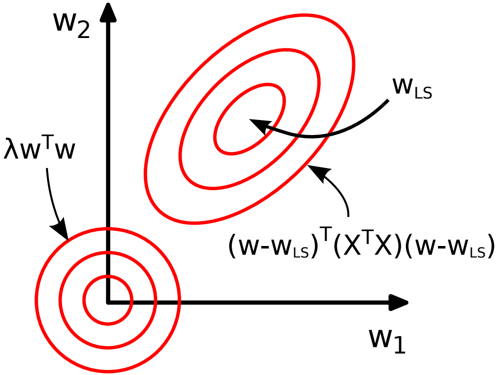
$$w_{\text{RR}} = (\lambda I + X^T X)^{-1} X^T y.$$

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.

RIDGE REGRESSION VS LEAST SQUARES I

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:

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

This gives a way of visualizing relationships.

We will discuss methods for picking λ later.