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

(extend by 1 dimension for the offset)

→ real-valued o/p

Data

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

what makes it a regression problem.

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.

(doesn't actually mean we have a linear function of the i/ps x.)

LEAST SQUARES (CONTINUED)

X → x_i along the rows
(n, d+1)

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

matrix   vector of errors

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

dot product btw i/p vector & coefficient w

dot product of error vector with itself we are summing the squares of those errors.

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

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 $\mathcal{L}$.

(find vector w at which gradient of this function is equal to 0.)

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?

PROBABILISTIC VIEW

Assuming that the covariance matrix is diagonal, we write covariance as sigma(variance) squared times an identity matrix.

Recall: Gaussian density in *n* dimensions (σ - variance)

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

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

output

y is random variable

formed by taking our i/ps and putting them along the rows of the matrix.

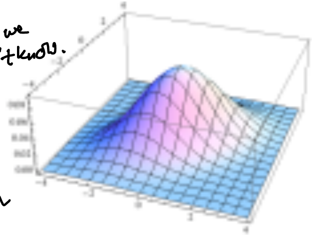
What if we restrict the mean to $\mu = Xw$ and find the *maximum likelihood* solution for w?

which we don't know.

X we know

w we don't know & we want to learn it.

μ is restricted to have this form



So we plug this matrix vector product (Xw) into the Gaussian (1) for μ. And then we say let's find the max. likelihood solⁿ to w.

Find the value of w that maximises the log-likelihood of the o/p vector y, given a mean this form, diagonal covariance, where each dimension has variance σ^2.

Maximum likelihood for Gaussian linear regression

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

$$w_{\text{ML}} = \arg\max_w \ \ln p(y|\mu = Xw, \sigma^2)$$

$$= \arg\max_w \ -\frac{1}{2\sigma^2}\|y - Xw\|^2 - \frac{n}{2}\ln(2\pi\sigma^2).$$

And now our goal is to find the vector w that maximizes this func. → which is a vector that maximizes the likelyhood of the data o/ps Y that we see given the corresponding i/ps X.

because This term doesn't involve w we don't even bother to write it

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

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

Intution: The least-square solution corresponds to the maximum likelihood sol^n of this multi-variate Gaussian assumption on our data.

PROBABILISTIC VIEW

So therefore in a sense what we can think of L.S. as doing is making an independent Gaussian noise assumption on the error. So this is some intuition that we can develop where we say that the L.S sol^n corresponds to the max. likelihood sol^n of multivariate Gaussian assumption on our data.

- 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:

Each of the errors are iid as 0 mean Gaussian with some σ^2. We are effectively modelling our o/p as being equal to the dot product of the i/p with weight w + iid Gaussian noise.

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?

this like an assumption we're going to make. We're going to say that our noise is i.i.d Gaussian, and that max. likelihood for this modelling assumption is the least squares solution.

* Another way to say, we're modelling our o/ps as independent Gaussian random variables where the mean is equal to the dot product btw the i/p x_i and coefficient vector w.

** As the previous slide we can simply say that we're making the assumption that vector Y is multivaried Gaussian with mean = Xw and covariance $\sigma^2 I$

i/p along each row

We're going to say that our noise is IID Gaussian. Essentially and that max. likelihood for the modelling assumption is the least-squared sol^n.

O/ps that we observe are generated Gaussian with mean Xw

Expected solution

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

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

nothing random there

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

we're making an distribution assumption on how y is generated - So y is random.

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

we don't know w. We are assuming, there exists some w.

This is the expectation function. So, of course every expectation you have some assumption of the distribution, you're taking that expectation over.
In this case, it's distribution on y. So you have a function of y times the distribution of y, and now we're integrating over y.

So what this is saying intuitively is that if we have
some ground truth value for W. And we generate,
we have some input, inputs X that we construct in a matrix
in this way.And then we generate an output Y according to
this distribution.
And then using that output y, we solve the maximum likelihood
solution for w. So we have the true w, and then using the random
vector y, solve the maximum likelihood solution for w, the expectation
of our maximum likelihood solution is equal to the truth.
This case what were saying is the maximum likelihood solution for
the vector w is an unbiased estimate.
Of the ground truth vector w, which we don't have access to.
So this is good, least squares or maximum likelihood for
this model is going to in expectation, give us the true parameter,
which is what we're trying to learn. I can't really understand the big thing, we just added a random variable with expectation equal to zero. And the expectation didn't change.

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?

If I have a Gaussian random variable with mean μ & variance σ^2. And σ^2 is huge, then even though that random variable in expectation is equal to the μ. Since the variance is so huge, I don't actually expect to see something close to the mean.

So, we have shown that maximum likelihood solⁿ is the correct one. But now we have calculate the variance of that solⁿ under the same modelling assumption.

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

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

covariance = expectation of the outer product with it's subtracted off.

[For Gaussian it is Σ.]

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

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

(expectation)

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

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

take the outer product

$E(y)=\mu$

1

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

Return to LS linear regression problem
And
Calculate the covariance of the max. likelihood sol^n under the Gaussian

Variance of the solution

Returning to least squares linear regression, we wish to find

max. likelihood sol^n – It's expectation

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

↳ From the previous slide where the variance of random vector is equal to the expectation of the outer product subtracted by the outer product of its mean. (1)

[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

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

→ expectation of the max. likelihood solⁿ is equal to the truth (already shown).

The sequence of equalities follows:[1]

$$\mathrm{Var}[w_{\mathrm{ML}}] = \mathbb{E}[\underbrace{(X^TX)^{-1}X^Ty}_{\text{max. likelihood sol}^n\text{ for } w}y^TX(X^TX)^{-1}] - ww^T$$

[whatever w is the expectation of ML is equal to w]

even though we don't know what it is. we can simply plug that value in thereas a placeholder.

[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]

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

metrices involving X are constant we can bring them out of the expectation.

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

PROBABILISTIC VIEW

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]

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

[we assume w has this relationship with o/p so Can plug this function in here.]

is equal to the covariance of y + outer product of mean of vector y which is equal to Xw.

[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]

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

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

PROBABILISTIC VIEW

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]

$$\begin{aligned}
\mathrm{Var}[w_{\mathrm{ML}}] &= \mathbb{E}[(X^TX)^{-1}X^Tyy^TX(X^TX)^{-1}] - ww^T \\
&= (X^TX)^{-1}X^T\mathbb{E}[yy^T]X(X^TX)^{-1} - ww^T \\
\sigma^2 (X^TX)^{-1} X^T \cancel{X} (\cancel{X^T} X)^{-1} \quad &= (X^TX)^{-1}X^T(\sigma^2 I + Xww^TX^T)X(X^TX)^{-1} - ww^T \\
= \sigma^2 (X^T X^{-1}) \longleftarrow \quad &= (X^TX)^{-1}X^T\sigma^2 IX(X^TX)^{-1} + \cdots \\
\mathrm{I} \quad \longleftarrow \quad & \underbrace{(X^TX)^{-1}X^TX}_{\mathrm{I}} ww^T \underbrace{X^TX(X^TX)^{-1}}_{\mathrm{I}} - ww^T \quad \Rightarrow ww^T - ww^T = 0 \\
&= \sigma^2(X^TX)^{-1}
\end{aligned}$$

---

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

Probabilistic view

If we make a Gaussian assumption, we assume that the vector of responses y is equal to matrix of feature vectors X times weight vector w + independent Gaussian noise σ^2.

Note: we have only defined a distribution of y, we haven't defined any distribution on x.

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

And we solve LS or max. likelihood sol^n for w, we assume there some true w. But we observe y which is a random vector acc. to $N(Xw, \sigma^2 I)$ distribution. We try to the inverse of learning what is w? Using max. likelihood.

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

- 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).

We don't trust max. likelihood sol^n when the values in covariance are large.

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

Then the expected value of our max. likelihood sol^n is equal to the true. But covariance is $\sigma^2 (X^T X)^{-1}$.

When the values in the covariance are v. v. large we can't say the max. likelihood sol^n is close to the truth.

Ridge Regression

REGULARIZED LEAST SQUARES

Why does penalizing the magnitude of w reduce the variance?

- We saw how with least squares, the values in w_{ML} may be huge.

and deviate wildly from the ground truth depending on what the matrix $(XX^T)^{-1}$ looks like.

- In general, when developing a model for data we often wish to *constrain* the model parameters in some way.

In this case, it might take the form of penalizing values of w that we might consider too large.

- There are many models of the form

penalty term

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

function of our model parameters w

sum of squared errors of our approximation

- The added terms are

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

↳ that penalizes values of the vector w in some way, in order to encourage properties of w that we might want in advance.

RIDGE REGRESSION

name of a model of a particular form, where we use a particular regularization function g.

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

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

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

g is squared magnitude of the vector w.

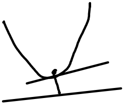
The term $g(w) = \|w\|^2$ penalizes large values in w.

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

- Case $\lambda \to 0$: $w_{\text{RR}} \to w_{\text{LS}}$ (II$^{\text{nd}}$ term disappears).
- Case $\lambda \to \infty$: $w_{\text{RR}} \to \vec{0}$ (any non-negative value of w has essentially infinite penalty.)

vector of all zeros.

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,

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

(In this case, we can't solve in closed form.)

Then, solve for w to find that

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

$\lambda \to 0$ least squares sol$^{\underline{n}}$

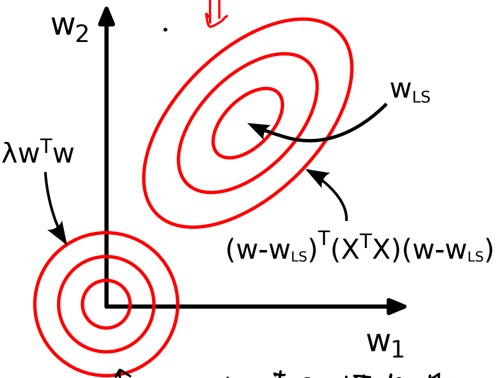
$\lambda \to \infty$, the inverse of that matrix is going to be a matrix of all zeros

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.

what z axis



same value at any pt along the circle

You can check that we can write:

$$\|\underline{y-Xw}\|^2+\lambda\|\underline{w}\|^2 = \overbrace{(w-w_{LS})^T(X^TX)(w-w_{LS})+\lambda w^T w}^{\text{involves } w}+\overbrace{(\text{const. w.r.t. } w)}^{\text{doesn't involve } w}.$$

write these 2 functions as their own independent level sets.

Means if you evaluate a function at any pt. along the curve, the set has the

DATA PREPROCESSING

[0-1, 1-100 diff. dimensions. Didn't have this problem with least squares, as we had put no restriction on w.]

penalizes each dimension of w equally. [$\lambda w^T w$]

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

[If two dimensions are equally important. Ridge regression is going to penalise on the dimension that takes smaller values.]

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.

why is it no needed?

Some Analysis of Ridge Regression

Ridge regression vs Least squares

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

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

Discuss what this λ is doing.

- 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 have more observations than dimensions in our problem

- We can write any $n \times d$ matrix X (assume $n > d$) as $X = USV^T$, where

  each column is unit length & orthogonal to every other column

  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. $\underbrace{V^T V = VV^T}_{\text{since it is square}} = I$. [Both columns & rows of V are orthonormal.]

     square

- From this we have the immediate equalities

  replacing the svd for X.

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

$VSU^T \; USV^T$

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

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

that's how come the inverse is $VS^{-2}V^T$

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

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

Using SVD to analyze the least squares solution.

Using the SVD we can rewrite the variance,

replace with svd

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

→ square, become closer to zero. then invert very large values.

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.)

↓ then we singular values that are very small.

The least squares prediction for new data is

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

→ very huge, for small values

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

depending upon how the vector x correlates with these singular vectors that small singular values.

Ridge regression vs Least squares I

Manipulating Ridge Regression sol^n in a certain way to relate it to the least squares sol^n.

Relationship to least squares solution

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

I (multiply & divide by the same thing.)

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

we can simply manipulate the least square sol^n acc. to this vector to get ridge regression sol^n? :)

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

We should be able to expect the ridge regression sol^n to have smaller magnitude than the least squares solution.

Adding something that's non-negative singular values to it. So you can imagine this as dividing by something that is >1. So you're shrinking this.

Ridge regression vs Least squares II

Replace x by its svd.

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

V → square orthogonal matrix
λ, S^{-2} → diagonal matrix

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

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

how?

I got this

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

$VMV^TVS^{-1}U^Ty \Rightarrow VMS^{-1}U^Ty$

As $S_{dd} \to 0$
① as $\lambda \to 0$ goes to $\to \infty$
② now $\lambda > 0$ goes to $\to 0$

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

Also, use svd on the solⁿ of least squares

[λ is essentially killing off these small singular values.]

RIDGE REGRESSION VS LEAST SQUARES III

how RR relates to LS as a LS problem.

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

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

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

d zeros

Attach a diagonal matrix along the bottom of X that has $\sqrt{\lambda}$ along the diagonal. AND solve the least square sol^n. for this problem

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.

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

So RR is almost like a augmented LS problem.

$\lambda \to \infty$, sol^n is a vector of 0s
$\lambda \to 0$, sol^n → LS sol^n

trace of sol^n as a function of λ.
way to understand how λ changes our rr sol^n.

plot of weight vector w. as a function of degrees of freedom.

Degrees of freedom:

$$df(\lambda) = \text{trace}\left[X(X^TX + \lambda I)^{-1}X^T\right]$$

why?

$$= \sum_{i=1}^{d} \frac{S_{ii}^2}{\lambda + S_{ii}^2}$$

Why is it equal to the trace of M?

This gives a way of visualizing relationships.

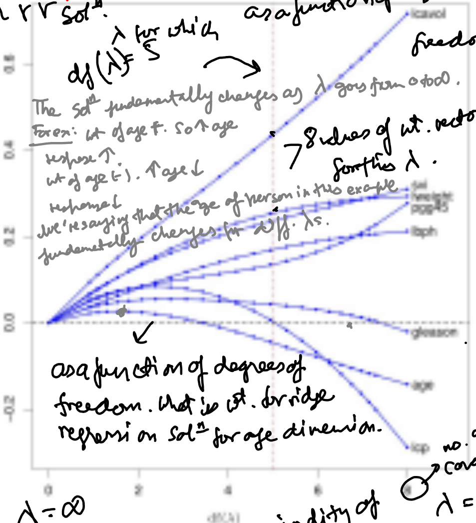
We will discuss methods for picking λ later.

how do you interpret this?
no. of df = dim. (# features)

df(λ) = λ for which S

The sol^n fundamentally changes as λ goes from 0 to ∞.
For ex: wt of age ↑. So ↑ age response ↑
wt of age (-). ↑ age ↓ response ↓
we're saying that the age of person in this example fundamentally changes for diff. λs.

→ 8 values of wt. vector for this λ.



as a function of degrees of freedom. What is wt. for ridge regression sol^n for age dimension.

λ = ∞

d is dimensionality of problem.

no. of covariates

λ = 0

df = d