ORIE 4741: Learning with Big Messy Data

Regularization

Professor Udell
Operations Research and Information Engineering
Cornell

September 10, 2019
Regularized empirical risk minimization

choose model by solving

\[
\text{minimize} \quad \sum_{i=1}^{n} \ell(x_i, y_i; w) + r(w)
\]

with variable \( w \in \mathbb{R}^d \)

- parameter vector \( w \in \mathbb{R}^d \)
- loss function \( \ell : \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^d \to \mathbb{R} \)
- regularizer \( r : \mathbb{R}^d \to \mathbb{R} \)
Regularized empirical risk minimization

choose model by solving

\[
\text{minimize } \sum_{i=1}^{n} \ell(x_i, y_i; w) + r(w)
\]

with variable \( w \in \mathbb{R}^d \)

- parameter vector \( w \in \mathbb{R}^d \)
- loss function \( \ell : \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^d \rightarrow \mathbb{R} \)
- regularizer \( r : \mathbb{R}^d \rightarrow \mathbb{R} \)

why?

- want to minimize the risk \( \mathbb{E}_{(x,y) \sim P} \ell(x, y; w) \)
- approximate it by the empirical risk \( \sum_{i=1}^{n} \ell(x, y; w) \)
- add regularizer to help model generalize
Example: regularized least squares

find best model by solving

$$\text{minimize} \quad \sum_{i=1}^{n} \ell(x_i, y_i; w) + r(w)$$

with variable $w \in \mathbb{R}^d$

ridge regression, aka quadratically regularized least squares:

- loss function $\ell(x, y; w) = (y - w^T x)^2$
- regularizer $r(w) = \|w\|^2$
Regularization

why regularize?

- reduce variance of the model
- impose prior structural knowledge
- improve interpretability
Regularization

why regularize?

▶ reduce variance of the model
▶ impose prior structural knowledge
▶ improve interpretability

why not regularize?

▶ Gauss-Markov theorem: least squares is the best linear unbiased estimator
▶ regularization adds bias
we might choose regularizer so models will be

- small
- sparse
- nonnegative
- smooth
- ...
we might choose regularizer so models will be

- small
- sparse
- nonnegative
- smooth
- ... 

comparison with forward- and backward-stepwise selection: 
regularized models tend to have lower variance.
see Elements of Statistical Learning (Hastie, Tibshirani, Friedman) for more information.
Quadratic regularizer

quadratic regularizer

\[ r(w) = \lambda \sum_{i=1}^{n} w_i^2 \]

ridge regression

\[
\text{minimize } \sum_{i=1}^{n} (y_i - w^T x_i)^2 + \lambda \sum_{i=1}^{n} w_i^2
\]

with variable \( w \in \mathbb{R}^d \)

solution \( w = (X^T X + \lambda I)^{-1} X^T y \)
Quadratic regularizer

- shrinks coefficients towards 0
- shrinks more in the direction of the smallest singular values of $X$
Is least squares scaling invariant?

suppose Alice and Bob do the same experiment

- Alice measures distance in mm
- Bob measures distance in km

they each compute an estimator with least squares and compare their predictions

Q: Do they make the same predictions?
A: Yes!
Is least squares scaling invariant?

suppose Alice and Bob do the same experiment

▶ Alice measures distance in mm
▶ Bob measures distance in km

they each compute an estimator with least squares and compare their predictions

Q: Do they make the same predictions?
Is least squares scaling invariant?

suppose Alice and Bob do the same experiment

- Alice measures distance in mm
- Bob measures distance in km

they each compute an estimator with least squares and compare their predictions

Q: Do they make the same predictions?
A: Yes!
Least squares is scaling invariant

if $\beta \in \mathbb{R}$, $D \in \mathbb{R}^{d \times d}$ is diagonal, and Alice’s measurements $(X', y')$ are related to Bob’s $(X, y)$ by

$$y' = \beta y, \quad X' = XD,$$

then the resulting least squares models are

$$w = (X^T X)^{-1} X^T y, \quad w' = (X'^T X')^{-1} X'^T y'$$

and they make the same predictions:

$$X'w' = X'(X'^T X')^{-1} X'^T y' = XD(D^T X^T XD)^{-1} D^T X^T \beta y$$

$$= XDD^{-1}(X^T X)^{-1}(D^T)^{-1} D^T X^T \beta y$$

$$= \beta X(X^T X)^{-1} X^T y = \beta Xw$$
Least squares is scaling invariant

if \( \beta \in \mathbb{R} \), \( D \in \mathbb{R}^{d \times d} \) is diagonal, and Alice’s measurements \((X', y')\) are related to Bob’s \((X, y)\) by

\[
y' = \beta y, \quad X' = XD,
\]

then the resulting least squares models are

\[
w = (X^T X)^{-1} X^T y, \quad w' = (X'^T X')^{-1} X'^T y'
\]

and they make the same predictions:

\[
X'w' = X'(X'^T X')^{-1} X'^T y' = XD(D^T X^T XD)^{-1} D^T X^T \beta y
\]
\[
= XDD^{-1}(X^T X)^{-1}(D^T)^{-1} D^T X^T \beta y
\]
\[
= \beta X(X^T X)^{-1} X^T y = \beta Xw
\]

we say least squares is **invariant under scaling**
Is ridge regression scaling invariant?

suppose Alice and Bob do the same experiment

▶ Alice measures distance in mm
▶ Bob measures distance in km

they each compute an estimator with ridge regression and compare their predictions
Is ridge regression scaling invariant?

suppose Alice and Bob do the same experiment

- Alice measures distance in mm
- Bob measures distance in km

they each compute an estimator with ridge regression and compare their predictions

Q: Do they make the same predictions?
Is ridge regression scaling invariant?

suppose Alice and Bob do the same experiment

- Alice measures distance in mm
- Bob measures distance in km

they each compute an estimator with ridge regression and compare their predictions

Q: Do they make the same predictions?

A: No!
Ridge regression is not scaling invariant

if $\beta \in \mathbb{R}$, $D \in \mathbb{R}^{d \times d}$ is diagonal, and Alice’s measurements $(X', y')$ are related to Bob’s $(X, y)$ by

$$y' = \beta y, \quad X' = XD,$$

then the resulting ridge regression models are

$$w = (X^T X + \lambda I)^{-1} X^T y, \quad w' = (X'^T X' + \lambda I)^{-1} X'^T y'$$

and the predictions are

$$Xw = X(X^T X + \lambda I)^{-1} X^T y, \quad X'w' = X'(X'^T X' + \lambda I)^{-1} X'^T y'$$

ridge regression is not invariant under coordinate transformations
Scaling and offsets

to get the same answer no matter the units of measurement, standardize the data: for each column of $X$ and of $y$

- demean: subtract column mean
- standardize: divide by column standard deviation

let

$$\mu_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij}, \quad \mu = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$\sigma_j^2 = \frac{1}{n} \sum_{i=1}^{n} (X_{ij} - \mu_j)^2, \quad \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mu)^2$$

solve

$$\text{minimize} \quad \sum_{i=1}^{n} \left( \frac{y_i - \mu}{\sigma} - \sum_{j=1}^{d} w_j \frac{X_{ij} - \mu_j}{\sigma_i} \right)^2 + \lambda \sum_{j=1}^{d} w_j^2$$
Scale the regularizer, not the data

instead of

\[
\text{minimize } \sum_{i=1}^{n} \left( \frac{y_i - \mu}{\sigma} - \sum_{j=1}^{d} w_j \frac{X_{ij} - \mu_i}{\sigma_i} \right)^2 + \sum_{j=1}^{d} w_j^2,
\]

- multiply through by \( \sigma^2 \)
- reparametrize \( w_j' = \frac{\sigma}{\sigma_j} w_j \)

to find the equivalent problem

\[
\text{minimize } \sum_{i=1}^{n} (y_i - \sum_{j=1}^{d} w_j' X_{ij} + c(w'))^2 + \sum_{j=1}^{d} \sigma_j^2 (w_j')^2,
\]

where \( c(w') \) is some linear function of \( w' \)
finally absorb \( c(w') \) into the constant term in the model

\[
\text{minimize } \|y - Xw'\|^2 + \lambda \sum_{j=1}^{d} \sigma_j^2 (w_j')^2,
\]
Scaling and offsets

a different solution to scaling and offsets: take the MAP view

- \( r(w) \) is negative log prior on \( w \)
- with a gaussian prior,
  \[
  r(w) = \sum_{i=1}^{n} \sigma_i^2 w_i^2
  \]
  where \( \frac{1}{\sigma_i} \) is the variance of the prior on the \( i \)th entry of \( w \)

- if you believe the noise in the \( i \)th features is large, penalize the \( i \)th entry more (\( \sigma_i \) big);
- if you believe the noise in the \( i \)th features is small, penalize the \( i \)th entry less (\( \sigma_i \) small);
- if you measure \( X \) or \( y \) in different units, your prior on \( w \) should change accordingly

example: don’t penalize the offset \( w_n \) of the model (\( \sigma_n \rightarrow \infty \))

\[
 r(w) = \sum_{i=1}^{n-1} w_i^2
\]
\( \ell_1 \) regularization

\( \ell_1 \) regularizer

\[ r(w) = \lambda \sum_{i=1}^{n} |w_i| \]

lasso problem

minimize

\[ \sum_{i=1}^{n} (y_i - w^T x_i)^2 + \lambda \sum_{i=1}^{n} |w_i| \]

with variable \( w \in \mathbb{R}^d \)

- penalizes large \( w \) less than ridge regression
- no closed form solution
Recall \( \ell_p \) norms

\( \ell_p \) norm \( \| w \|_p \) for \( p \in (0, \infty) \) is defined as

\[
\| w \|_p = \left( \sum_{i=1}^{d} |w|^p \right)^{1/p}
\]

Examples:

- \( \ell_1 \) norm is \( \| w \|_1 = \sum_{i=1}^{d} |w| \)
- \( \ell_2 \) norm is \( \| w \|_2 = \sqrt{\sum_{i=1}^{d} w^2} \)

For \( p = 0 \) or \( p = \infty \), \( \ell_p \) norm is defined by taking limit:

- \( \ell_\infty \) norm is \( \| w \|_\infty = \lim_{p \to \infty} \left( \sum_{i=1}^{d} |w|^p \right)^{1/p} = \max_i |w_i| \)
- \( \ell_0 \) norm is \( \| w \|_0 = \lim_{p \to 0} \left( \sum_{i=1}^{d} |w|^p \right)^{1/p} = \text{card}(w) \), number of nonzeros in \( w \)

Note: \( \ell_0 \) is not actually a norm

(not absolutely homogeneous since \( \| \alpha w \|_0 = \| w \|_0 \) for \( \alpha \neq 0 \))
\( \ell_1 \) regularization

why use \( \ell_1 \)?

▶ best convex lower bound for \( \ell_0 \) on the \( \ell_\infty \) unit ball
▶ tends to produce sparse solution

example:

▶ suppose \( X_{:1} = y, \ X_{:2} = \alpha y \) for some \( \alpha > 0 \)
▶ fit lasso model and ridge regression model as \( \lambda \to 0 \)

\[
\begin{align*}
\hat{w}^{\text{ridge}} &= \lim_{\lambda \to 0} \text{argmin}_{w} \|y - Xw\|^2 + \lambda \|w\|^2_2 \\
\hat{w}^{\text{lasso}} &= \lim_{\lambda \to 0} \text{argmin}_{w} \|y - Xw\|^2 + \lambda \|w\|^1_1
\end{align*}
\]

▶ as \( \lambda \to 0 \), solution has \( w_1 + \alpha w_2 = 1 \)
$\ell_1$ regularization

why use $\ell_1$?

▶ best convex lower bound for $\ell_0$ on the $\ell_\infty$ unit ball
▶ tends to produce sparse solution

dexample:

▶ suppose $X_1 = y$, $X_2 = \alpha y$ for some $\alpha > 0$
▶ fit lasso model and ridge regression model as $\lambda \to 0$

$$w^{\text{ridge}} = \lim_{\lambda \to 0} \arg\min_w \|y - Xw\|^2 + \lambda \|w\|_2^2$$

$$w^{\text{lasso}} = \lim_{\lambda \to 0} \arg\min_w \|y - Xw\|^2 + \lambda \|w\|_1$$

▶ as $\lambda \to 0$, solution has $w_1 + \alpha w_2 = 1$
▶ ridge regression minimizes $w_1^2 + w_2^2 \implies w_1 = w_2 = \frac{1}{2}$
▶ lasso minimizes $|w_1| + |w_2| \implies w_1 = 1, w_2 = 0$ is valid
**Sparsity**

why would you want sparsity?

- credit card application: requires less info from applicant
- medical diagnosis: easier to explain model to doctor
- genomic study: which genes to investigate?
**Convex indicator**

define **convex indicator** $1 : \{\text{true, false}\} \rightarrow \mathbb{R} \cup \{\infty\}$

$$1(z) = \begin{cases} 
0 & z \text{ is true} \\
\infty & z \text{ is false}
\end{cases}$$

define **convex indicator** of set $C$

$$1_C(x) = 1(x \in C) = \begin{cases} 
0 & x \in C \\
\infty & \text{otherwise}
\end{cases}$$
**Convex indicator**

Define **convex indicator** $1 : \{\text{true, false}\} \rightarrow \mathbb{R} \cup \{\infty\}$

$$1(z) = \begin{cases} 
0 & \text{if } z \text{ is true} \\
\infty & \text{if } z \text{ is false}
\end{cases}$$

Define **convex indicator** of set $C$

$$1_C(x) = 1(x \in C) = \begin{cases} 
0 & x \in C \\
\infty & \text{otherwise}
\end{cases}$$

don’t confuse this with the boolean indicator $\mathbb{1}(z)$ (no standard notation... )
Nonnegative regularization

nonnegative regularizer

\[ r(w) = \sum_{i=1}^{n} 1(w_i \geq 0) \]

nonnegative least squares problem (NNLS)

\[
\text{minimize } \sum_{i=1}^{n} (y_i - w^T x_i)^2 + \lambda \sum_{i=1}^{n} 1(w_i \geq 0)
\]

with variable \( w \in \mathbb{R}^d \)

- value is \( \infty \) if \( w_i < 0 \)
- so solution is always nonnegative
- often, solution is also sparse
Nonnegative coefficients

why would you want nonnegativity?
Nonnegative coefficients

why would you want nonnegativity?

- electricity usage: how often is device turned on?
  - n = times, d = electric devices,
  - y = usage, X = which devices use power at which times
  - w = devices used by household

- hyperspectral imaging: which species are present?
  - n = frequencies, d = possible materials,
  - y = observed spectrum, X = known spectrum of each material
  - w = material composition of location

- logistics: which routes to run?
  - n = locations, d = possible routes,
  - y = demand, X = which routes visit which locations
  - w = size of truck to send on each route
Nonnegative coefficients

why would you want nonnegativity?

▶ electricity usage: how often is device turned on?
  ▶ n = times, d = electric devices,
  ▶ y = usage, X = which devices use power at which times
  ▶ w = devices used by household

▶ hyperspectral imaging: which species are present?
  ▶ n = frequencies, d = possible materials,
  ▶ y = observed spectrum, X = known spectrum of each material
  ▶ w = material composition of location
Nonnegative coefficients

why would you want nonnegativity?

- electricity usage: how often is device turned on?
  - n = times, d = electric devices,
  - y = usage, X = which devices use power at which times
  - w = devices used by household

- hyperspectral imaging: which species are present?
  - n = frequencies, d = possible materials,
  - y = observed spectrum, X = known spectrum of each material
  - w = material composition of location

- logistics: which routes to run?
  - n = locations, d = possible routes,
  - y = demand, X = which routes visit which locations
  - w = size of truck to send on each route
Demo: Regularized Regression

https://github.com/ORIE4741/demos/
RegularizedRegression.ipynb
Smooth coefficients

smooth regularizer

\[ r(w) = \sum_{i=1}^{d-1} (w_{i+1} - w_i)^2 = \|Dw\|^2 \]

where \( D \in \mathbb{R}^{(d-1) \times d} \) is the first order difference operator

\[ D_{ij} = \begin{cases} 
1 & j = i \\
-1 & j = i + 1 \\
0 & \text{else} 
\end{cases} \]

smoothed least squares problem

\[
\text{minimize} \quad \sum_{i=1}^{n} (y_i - w^T x_i)^2 + \lambda \|Dw\|^2
\]
Why smooth?

- allow model to change over space or time
  - e.g., different years in tax data
- interpolates between one model and separate models for different domains
  - e.g., counties in tax data
- can couple any pairs of model coefficients, not just \((i, i + 1)\)