ORIE 4741: Learning with Big Messy Data

Regularization

Professor Udell

Operations Research and Information Engineering
Cornell

October 28, 2021

Announcements 10/26/21

- hw4 out, due 10am 11/1
  - save slip days for emergencies
- project midterm report due 11:59pm 11/1
- section this week: optimization algorithms for regularized problems

Announcements 10/28/21

- hw4 out, due 10am 11/1
  - save slip days for emergencies
  - talk with me if you run out of slip days
  - turn in hw early, then have fun on Halloween!
- project midterm report due 11:59pm 11/1
  - your peers are grading you; make your report make sense to them
  - look at previous years reports for organizational ideas
  - "three techniques from class": look ahead in the course topics and/or ask
  - look at the peer grading rubric (on projects webpage)

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 $w \in \mathbf{R}^d$

- parameter vector $w \in \mathbf{R}^d$
- loss function $\ell : \mathcal{X} \times \mathcal{Y} \times \mathbf{R}^d \to \mathbf{R}$
- regularizer $r : \mathbf{R}^d \to \mathbf{R}$

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 $w \in \mathbf{R}^d$

- parameter vector $w \in \mathbf{R}^d$
- loss function $\ell : \mathcal{X} \times \mathcal{Y} \times \mathbf{R}^d \to \mathbf{R}$
- regularizer $r : \mathbf{R}^d \to \mathbf{R}$

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 $w \in \mathbf{R}^d$

ridge regression, aka quadratically regularized least squares:

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

Outline

Regularizers

ℓ_1 regularizization

ControlBurn: Ensembles + Lasso

Nonnegative regularizer

Quadratic regularizization

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 increases bias

Regularizers: a tour

we might choose regularizer so models will be

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

Regularizers: a tour

we might choose regularizer so models will be

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

compared with forward- and backward-stepwise selection (*e.g.*, AIC, BIC), regularized models tend to have **lower variance**.

source: Elements of Statistical Learning (Hastie, Tibshirani, Friedman)

Outline

Regularizers

ℓ_1 regularizization

ControlBurn: Ensembles + Lasso

Nonnegative regularizer

Quadratic regularizization

ℓ_1 regularization

ℓ_1 regularizer

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

ℓ_1 regularization

ℓ_1 regularizer

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

lasso problem

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

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

ℓ_1 regularization

ℓ_1 regularizer

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

lasso problem

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

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

- penalizes large w less than quadratic regularization
- no closed form solution

Recall ℓ_p norms

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

$$\|w\|_p = (\sum_{i=1}^{d} |w|^p)^{1/p}$$

Recall ℓ_p norms

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

$$\|w\|_p = (\sum_{i=1}^{d} |w|^p)^{1/p}$$

examples:

Recall ℓ_p norms

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

$$\|w\|_p = (\sum_{i=1}^{d} |w|^p)^{1/p}$$

examples:

- ℓ_1 norm is $\|w\|_1 = \sum_{i=1}^{d} |w|$

Recall ℓ_p norms

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

$$\|w\|_p = (\sum_{i=1}^{d} |w|^p)^{1/p}$$

examples:

- ℓ_1 norm is $\|w\|_1 = \sum_{i=1}^{d} |w|$
- ℓ_2 norm is $\|w\|_2 = \sqrt{\sum_{i=1}^{d} w^2}$

Recall ℓ_p norms

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

$$\|w\|_p = (\sum_{i=1}^{d} |w|^p)^{1/p}$$

examples:

- ℓ_1 norm is $\|w\|_1 = \sum_{i=1}^{d} |w|$
- ℓ_2 norm is $\|w\|_2 = \sqrt{\sum_{i=1}^{d} w^2}$

for $p = 0$ or $p = \infty$, ℓ_p norm is defined by taking limit:

Recall ℓ_p norms

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

$$\|w\|_p = (\sum_{i=1}^{d} |w|^p)^{1/p}$$

examples:

- ℓ_1 norm is $\|w\|_1 = \sum_{i=1}^{d} |w|$
- ℓ_2 norm is $\|w\|_2 = \sqrt{\sum_{i=1}^{d} w^2}$

for $p = 0$ or $p = \infty$, ℓ_p norm is defined by taking limit:

- ℓ_∞ norm is $\|w\|_\infty = \lim_{p\to\infty} (\sum_{i=1}^{d} |w|^p)^{1/p} = \max_i |w_i|$

Recall ℓ_p norms

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

$$\|w\|_p = (\sum_{i=1}^{d} |w|^p)^{1/p}$$

examples:

- ℓ_1 norm is $\|w\|_1 = \sum_{i=1}^{d} |w|$
- ℓ_2 norm is $\|w\|_2 = \sqrt{\sum_{i=1}^{d} w^2}$

for $p = 0$ or $p = \infty$, ℓ_p norm is defined by taking limit:

- ℓ_∞ norm is $\|w\|_\infty = \lim_{p\to\infty}(\sum_{i=1}^{d} |w|^p)^{1/p} = \max_i |w_i|$
- ℓ_0 norm is $\|w\|_0 = \lim_{p\to 0}(\sum_{i=1}^{d} |w|^p)^{1/p} = \mathbf{card}(w)$, number of nonzeros in w

Recall ℓ_p norms

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

$$\|w\|_p = (\sum_{i=1}^{d} |w|^p)^{1/p}$$

examples:

- ℓ_1 norm is $\|w\|_1 = \sum_{i=1}^{d} |w|$
- ℓ_2 norm is $\|w\|_2 = \sqrt{\sum_{i=1}^{d} w^2}$

for $p = 0$ or $p = \infty$, ℓ_p norm is defined by taking limit:

- ℓ_∞ norm is $\|w\|_\infty = \lim_{p \to \infty}(\sum_{i=1}^{d} |w|^p)^{1/p} = \max_i |w_i|$
- ℓ_0 norm is $\|w\|_0 = \lim_{p \to 0}(\sum_{i=1}^{d} |w|^p)^{1/p} = \mathbf{card}(w)$, number of nonzeros in w

technical note: ℓ_0 is not actually a norm
(not absolutely homogeneous since $\|\alpha w\|_0 = \|w\|_0$ for $\alpha \neq 0$)

ℓ_1 regularization

why use ℓ_1?

- best convex lower bound for ℓ_0 on the ℓ_∞ unit ball
- tends to produce sparse solution

ℓ_1 vs ℓ_2 regularization

- suppose two features, same up to scaling: $X_{:1} = y$, $X_{:2} = y$
- fit lasso model and ridge regression model as $\lambda \to 0$

$$\begin{aligned} w^{\text{ridge}} &= \lim_{\lambda \to 0} \operatorname*{argmin}_w \|y - Xw\|^2 + \lambda \|w\|_2^2 \\ w^{\text{lasso}} &= \lim_{\lambda \to 0} \operatorname*{argmin}_w \|y - Xw\|^2 + \lambda \|w\|_1 \end{aligned}$$

- as $\lambda \to 0$, solution solves least squares $\implies w_1 + w_2 = 1$

ℓ_1 vs ℓ_2 regularization

- suppose two features, same up to scaling: $X_{:1} = y$, $X_{:2} = y$
- fit lasso model and ridge regression model as $\lambda \to 0$

$$
\begin{aligned}
w^{\text{ridge}} &= \lim_{\lambda \to 0} \operatorname*{argmin}_{w} \|y - Xw\|^2 + \lambda \|w\|_2^2 \\
w^{\text{lasso}} &= \lim_{\lambda \to 0} \operatorname*{argmin}_{w} \|y - Xw\|^2 + \lambda \|w\|_1
\end{aligned}
$$

- as $\lambda \to 0$, solution solves least squares $\implies w_1 + w_2 = 1$
- quadratic regularization minimizes $w_1^2 + w_2^2 \implies$
  - A. $w_1 = w_2 = \frac{1}{2}$
  - B. $w_1 = 1,\ w_2 = 0$
  - C. $w_1 = 0,\ w_2 = 1$

ℓ_1 vs ℓ_2 regularization

- suppose two features, same up to scaling: $X_{:1} = y$, $X_{:2} = y$
- fit lasso model and ridge regression model as $\lambda \to 0$

$$
\begin{aligned}
w^{\text{ridge}} &= \lim_{\lambda \to 0} \operatorname*{argmin}_w \|y - Xw\|^2 + \lambda \|w\|_2^2 \\
w^{\text{lasso}} &= \lim_{\lambda \to 0} \operatorname*{argmin}_w \|y - Xw\|^2 + \lambda \|w\|_1
\end{aligned}
$$

- as $\lambda \to 0$, solution solves least squares $\implies$ $w_1 + w_2 = 1$
- quadratic regularization minimizes $w_1^2 + w_2^2 \implies$
  - A. $w_1 = w_2 = \frac{1}{2}$
  - B. $w_1 = 1,\ w_2 = 0$
  - C. $w_1 = 0,\ w_2 = 1$

  $w_1 = w_2 = \frac{1}{2}$

ℓ_1 vs ℓ_2 regularization

- suppose two features, both equal: $X_{:1} = y$, $X_{:2} = y$
- fit lasso model and ridge regression model as $\lambda \to 0$

$$
\begin{aligned}
w^{\text{ridge}} &= \lim_{\lambda \to 0} \operatorname*{argmin}_w \|y - Xw\|^2 + \lambda \|w\|_2^2 \\
w^{\text{lasso}} &= \lim_{\lambda \to 0} \operatorname*{argmin}_w \|y - Xw\|^2 + \lambda \|w\|_1
\end{aligned}
$$

- as $\lambda \to 0$, solution solves least squares $\implies w_1 + w_2 = 1$

ℓ_1 vs ℓ_2 regularization

- suppose two features, both equal: $X_{:1} = y$, $X_{:2} = y$
- fit lasso model and ridge regression model as $\lambda \to 0$

$$
\begin{aligned}
w^{\text{ridge}} &= \lim_{\lambda \to 0} \operatorname*{argmin}_w \|y - Xw\|^2 + \lambda \|w\|_2^2 \\
w^{\text{lasso}} &= \lim_{\lambda \to 0} \operatorname*{argmin}_w \|y - Xw\|^2 + \lambda \|w\|_1
\end{aligned}
$$

- as $\lambda \to 0$, solution solves least squares $\implies w_1 + w_2 = 1$
- lasso minimizes $|w_1| + |w_2| \implies$
  A. $w_1 = w_2 = \frac{1}{2}$
  B. $w_1 = 1,\ w_2 = 0$
  C. $w_1 = 0,\ w_2 = 1$

ℓ_1 vs ℓ_2 regularization

- suppose two features, both equal: $X_{:1} = y$, $X_{:2} = y$
- fit lasso model and ridge regression model as $\lambda \to 0$

$$
\begin{aligned}
w^{\text{ridge}} &= \lim_{\lambda \to 0} \operatorname*{argmin}_w \|y - Xw\|^2 + \lambda \|w\|_2^2 \\
w^{\text{lasso}} &= \lim_{\lambda \to 0} \operatorname*{argmin}_w \|y - Xw\|^2 + \lambda \|w\|_1
\end{aligned}
$$

- as $\lambda \to 0$, solution solves least squares $\implies w_1 + w_2 = 1$
- lasso minimizes $|w_1| + |w_2| \implies$

  A. $w_1 = w_2 = \frac{1}{2}$
  B. $w_1 = 1,\ w_2 = 0$
  C. $w_1 = 0,\ w_2 = 1$

  all options are equally good

ℓ_1 vs ℓ_2 regularization

- suppose two features, same up to scaling $0 < \alpha < 1$: $X_{:1} = y$, $X_{:2} = \alpha y$
- fit lasso model and ridge regression model as $\lambda \to 0$

$$
\begin{aligned}
w^{\text{ridge}} &= \lim_{\lambda \to 0} \operatorname*{argmin}_{w} \|y - Xw\|^2 + \lambda \|w\|_2^2 \\
w^{\text{lasso}} &= \lim_{\lambda \to 0} \operatorname*{argmin}_{w} \|y - Xw\|^2 + \lambda \|w\|_1
\end{aligned}
$$

- as $\lambda \to 0$, solution solves least squares $\implies w_1 + \alpha w_2 = 1$

ℓ_1 vs ℓ_2 regularization

- suppose two features, same up to scaling $0 < \alpha < 1$: $X_{:1} = y$, $X_{:2} = \alpha y$
- fit lasso model and ridge regression model as $\lambda \to 0$

$$
\begin{aligned}
w^{\text{ridge}} &= \lim_{\lambda \to 0} \operatorname*{argmin}_w \|y - Xw\|^2 + \lambda \|w\|_2^2 \\
w^{\text{lasso}} &= \lim_{\lambda \to 0} \operatorname*{argmin}_w \|y - Xw\|^2 + \lambda \|w\|_1
\end{aligned}
$$

- as $\lambda \to 0$, solution solves least squares $\implies w_1 + \alpha w_2 = 1$
- lasso minimizes $|w_1| + |w_2| \implies$
  - A. $w_1 = 1/2$, $w_2 = 1/2\alpha$
  - B. $w_1 = 1$, $w_2 = 0$
  - C. $w_1 = 0$, $w_2 = 1/\alpha$

ℓ_1 vs ℓ_2 regularization

- suppose two features, same up to scaling $0 < \alpha < 1$: $X_{:1} = y$, $X_{:2} = \alpha y$
- fit lasso model and ridge regression model as $\lambda \to 0$

$$
\begin{aligned}
w^{\text{ridge}} &= \lim_{\lambda \to 0} \operatorname*{argmin}_w \|y - Xw\|^2 + \lambda \|w\|_2^2 \\
w^{\text{lasso}} &= \lim_{\lambda \to 0} \operatorname*{argmin}_w \|y - Xw\|^2 + \lambda \|w\|_1
\end{aligned}
$$

- as $\lambda \to 0$, solution solves least squares $\implies w_1 + \alpha w_2 = 1$
- lasso minimizes $|w_1| + |w_2| \implies$
  - A. $w_1 = 1/2$, $w_2 = 1/2\alpha$
  - B. $w_1 = 1$, $w_2 = 0$
  - C. $w_1 = 0$, $w_2 = 1/\alpha$

  $w_1 = 1$, $w_2 = 0$

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?

Outline

ControlBurn

paper: https://arxiv.org/abs/2107.00219

demo: https://github.com/udellgroup/controlburn/blob/main/Demo/ControlBurnDemoNotebook.ipynb

Outline

Regularizers

ℓ_1 regularizization

ControlBurn: Ensembles + Lasso

Nonnegative regularizer

Quadratic regularizization

Convex indicator

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

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

define **convex indicator** of set C

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

Convex indicator

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

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

define **convex indicator** of set C

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

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

Nonnegative regularization

nonnegative regularizer

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

nonnegative least squares problem (NNLS)

$$\text{minimize} \quad \sum_{i=1}^{n} (y_i - w^T x_i)^2 + \sum_{i=1}^{n} \mathbf{1}(w_i \geq 0)$$

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

- value is ∞ 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

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

Outline

Quadratic regularizer

quadratic regularizer

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

ridge regression

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

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

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

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

B. no

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

B. no

A: Yes!

Least squares is scaling invariant

if $\beta \in \mathbf{R}$, $D \in \mathbf{R}^{d \times d}$ 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:

$$\begin{aligned} X'w' &= X'(X'^T X')^{-1} X'^T y' = XD(D^T X^T X D)^{-1} D^T X^T \beta y \\ &= XDD^{-1}(X^T X)^{-1}(D^T)^{-1} D^T X^T \beta y \\ &= \beta X (X^T X)^{-1} X^T y = \beta X w \end{aligned}$$

Least squares is scaling invariant

if $\beta \in \mathbf{R}$, $D \in \mathbf{R}^{d \times d}$ 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:

$$\begin{aligned} X'w' &= X'(X'^T X')^{-1} X'^T y' = XD(D^T X^T X D)^{-1} D^T X^T \beta y \\ &= XDD^{-1}(X^T X)^{-1}(D^T)^{-1} D^T X^T \beta y \\ &= \beta X (X^T X)^{-1} X^T y = \beta X w \end{aligned}$$

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?

A. yes

B. no

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. yes

B. no

A: No!

Ridge regression is not scaling invariant

if $\beta \in \mathbf{R}$, $D \in \mathbf{R}^{d \times d}$ 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

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

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

solve

$$\text{minimize} \quad \sum_{i=1}^{n}\left(\frac{y_i - \mu}{\sigma} - \sum_{j=1}^{d} w_j \frac{X_{ij} - \mu_j}{\sigma_j}\right)^2 + \lambda \sum_{j=1}^{d} w_j^2$$

Scale the regularizer, not the data

instead of

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

- multiply through by σ^2
- reparametrize $w'_j = \frac{\sigma}{\sigma_j} w_j$

to find the equivalent problem

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

where $c(w')$ is some linear function of w'

finally absorb $c(w')$ into the constant term in the model

$$\text{minimize} \quad \|y - Xw'\|^2 + \lambda \sum_{j=1}^{d} \sigma_j^2 (w'_j)^2,$$

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

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 ith entry of w
- if you believe the noise in the ith features is large, penalize the ith entry more (σ_i big);
- if you believe the noise in the ith features is small, penalize the ith entry less (σ_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 ($\sigma_n \to \infty$)

$$r(w) = \sum^{n-1} w_i^2$$

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 $D \in \mathbf{R}^{(d-1)\times d}$ 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 $(i, i+1)$