# Normal Equation and Least Squares

The least Squares problem:

$$\min_{\beta} \ \ell(\beta) = \| y - X\beta \|_2^2 , \qquad X \in R^{n\times p} \text{ data matrix}$$

$$= y^Ty - 2y^TX\beta + \beta^TX^TX\beta \qquad y \in R^{n\times 1} \text{ target values}$$

$n$: number of sample points

$p$: dimension of feature vectors.

Solve by setting the gradient to zero:

$$\nabla_\beta \ell(\beta) = -2X^Ty + 2X^TX\beta = 0$$

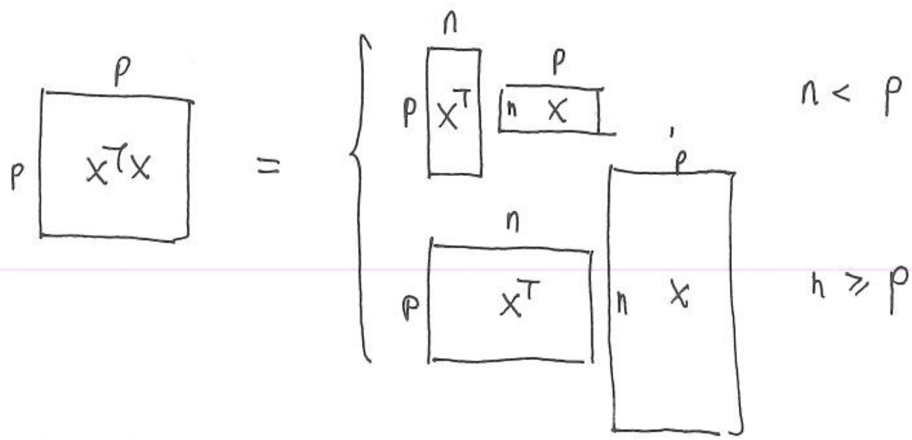
$$\Leftrightarrow \quad X^TX\beta = X^Ty , \quad \text{called the "normal equation."}$$

If $X^TX$ is invertible, $\hat{\beta} = (X^TX)^{-1}X^Ty$ is the unique solution.

Q: Under what condition is $(X^TX)$ invertible, or equivalent, of full rank?

Note: The rank of a square matrix is the max # of linearly independent rows (or columns).

A: Two cases: ① $n < p$ ② $n \geq p$.




$\text{rank}(X^TX) \leq n$,

$\Rightarrow$ because every column of $X^TX$ is a linear combination of at most $n$ $p$-dimensional vectors.

When $n < p$, $\text{rank}(X^TX) < p$, so $X^TX$ not invertible, and the least square problem has multiple solutions.

When $n \geq p$, and there are $p$ linearly independent feature vectors in the data, (which is usually the case when $n > p$), $X^TX$ is invertible and

$\hat{\beta} = (X^TX)^{-1}X^Ty$ is the unique solution.

# Ridge Regression

$$\min_{\beta} \ell_{ridge}(\beta) = \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$$

$$= y^Ty - 2y^TX\beta + \beta^T(X^TX + \lambda I)\beta$$

$\lambda > 0$ : regularization parameter,

$I$ : p-by-p identity matrix

Solve $\nabla_\beta \ell_{ridge}(\beta) = 0$

$$\Leftrightarrow \quad -2X^Ty + 2(X^TX + \lambda I)\beta = 0$$

$$\Leftrightarrow \quad (X^TX + \lambda I)\beta = X^Ty.$$

Thm: $X^TX + \lambda I$ is always invertible

pf: Prove the following lemma first:

> Lemma: $\forall\, a \in \mathbb{R}^p$, $a$ not the zero vector,
>
> $$a^T(X^TX + \lambda I)a > 0.$$
>
> pf: $a^T(X^TX + \lambda I)a = a^TX^TXa + \lambda a^Ta$
>
> $$= \|Xa\|_2^2 + \lambda a^Ta > 0.$$ since $a \neq 0$ and $\lambda > 0$

Then prove by contradiction: If $X^TX + \lambda I$ is not invertible, its columns are not linearly independent, so there exists $a \in \mathbb{R}^p$, $a \neq 0$ such that

$$(X^TX + \lambda I)a = 0,$$

which implies $a^T(X^TX + \lambda I)a = 0$, a contradiction to the lemma.

#

So

$$\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T y$$

is the unique solution to the Ridge Regression problem.

Why ridge regression?

① When $n < p$, helps to get a unique solution.

② When $n \geq p$, even though $\hat{\beta}$ usually exists and is unique, it may overfit the data. In terms of Bias and variance,

$\text{bias}(\hat{\beta}_{ridge}) > \text{bias}(\hat{\beta}) = 0$ under the linear model,

$\text{Variance}(\hat{\beta}_{ridge}) < \text{variance}(\hat{\beta})$

As $\lambda \uparrow$, $\text{bias}(\hat{\beta}_{ridge}) \uparrow$ and $\text{Variance}(\hat{\beta}_{ridge}) \downarrow$

Use cross validation to decide $\lambda$.

# Histogram.

Consider the following family of p.d.f.s over the 1-d interval $[a,b]$:

$$f(x) = \sum_{j=1}^{k} \mathbb{1}\{x \in Bin_j\}\, p_j \quad , \quad p_j \geq 0 \text{ is the density in the } j\text{th bin.}$$

Let $\Delta_1, \Delta_2, \ldots, \Delta_k$ be the pre-specified sizes of the $k$ bins, so $\sum_{j=1}^{k} \Delta_j = b-a$

and $\text{Prob}(X \in Bin_j) = \int_a^b \mathbb{1}\{x \in Bin_j\} f(x)\, dx = p_j \Delta_j$.

Since $f(x)$ is a p.d.f. we have

$$\int_a^b f(x)\, dx = \sum_{j=1}^{k} p_j \Delta_j = 1$$

Given an i.i.d sample $\{x_1, x_2, \ldots x_n\}$ drawn from some $f$ in this family, we want to estimate the densities $p_1, p_2, \ldots, p_k$. We do ML estimation.

Likelihood: $L(p_1 \ldots p_k) = \prod_{i=1}^{n} \prod_{j=1}^{k} (p_j \Delta_j)^{\mathbb{1}\{x_i \in Bin_j\}}$

Log likelihood: $\ell(p_1 \ldots p_k) = \sum_{i=1}^{n} \sum_{j=1}^{k} \mathbb{1}\{x_i \in Bin_j\} \log(p_j \Delta_j)$

$$= \sum_{j=1}^{k} \underbrace{\sum_{i=1}^{n} \mathbb{1}\{x_i \in Bin_j\}}_{\text{call } n_j, \text{ \# of points in } Bin_j} \log(p_j \Delta_j)$$

concave in $p_1 \ldots p_k$

Solve $\max \ell(p_1, \ldots, p_k)$ s.t. $\sum_j p_j \Delta_j = 1$ by setting the gradient of the Lagrangian function to zero:

$$\partial_{p_{j'}} \left[ \ell(p_1 \ldots p_k) - \lambda \left( \sum_j p_j \Delta_j - 1 \right) \right] = 0 \iff \frac{n_{j'}}{p_{j'}} - \lambda \Delta_{j'} = 0$$

$\therefore p_{j'} = \dfrac{n_{j'}}{\lambda \Delta_{j'}}$. Since $\sum_{j'} p_{j'} \Delta_{j'} = 1$, $\lambda$ must be $\sum_{j'} n_{j'} = n$, and

$p_{j'} = \dfrac{n_{j'}}{n \Delta_{j'}}$, the histogram density estimate.

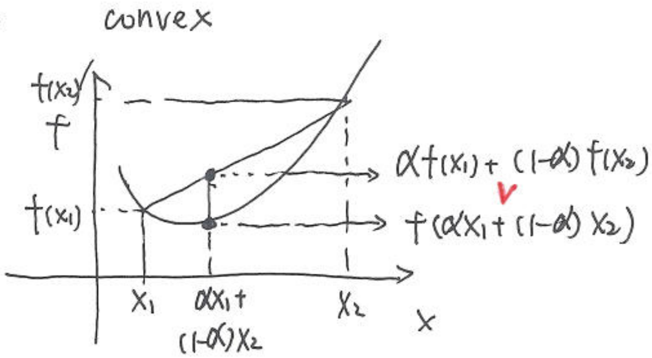
# $L_0$ penalty is non-convex

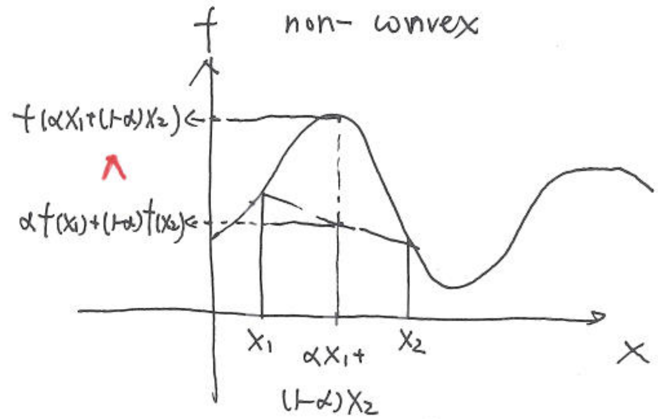
For simplicity, consider one dimensional case.

Def. A function $f$ is convex if

$$\alpha f(x_1) + (1-\alpha) f(x_2) \geq f(\alpha x_1 + (1-\alpha) x_2) \quad \forall\ 0 \leq \alpha \leq 1 \text{ and } \forall\ x_1, x_2 \text{ in domain of } f.$$

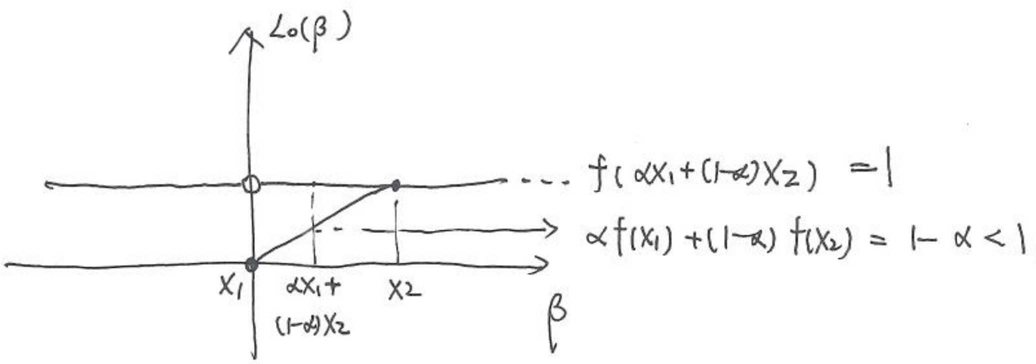
Ex.




The $L_0$ penalty in 1-d:

$$L_0(\beta) = \mathbb{1}\{\beta \neq 0\}$$




$\Rightarrow$ non-convex.