ML and Opt Lecture 1, 8/29/23

- Logistics
- What is Optimization?
- What is ML? *
- Example of a ML model: logistic regression *
- Example of gradient descent vs a specialized algorithm for a convex optim prob: linear regression *

* postponed to next lecture

* code available on website

Optimization

Given an objective function J, our goal is to either

$$x^* = \underset{x}{\operatorname{argmax}}\ J(x) \qquad x^* = \underset{x}{\operatorname{argmin}}\ J(x)$$

(unconstrained optimization)

the set of x such that $J(x) \geq J(y)$ for any other choices y

the set of x such that $J(x) \leq J(y)$ for any other choices y

$$x^* = \underset{x \in C}{\operatorname{argmax}}\ J(x) \qquad x^* = \underset{x \in C}{\operatorname{argmin}}\ J(x)$$

(constrained optimization)

here C is a set of x defined by some constraints

Ex: Solving a least-squares problem

$$\begin{array}{cc} & d \\ n & \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \end{array} \qquad \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad Y \in \mathbb{R}^n$$

X – collection of data points in $\mathbb{R}^d$

Y – corresponding observations

$\|x\|_2 = \sqrt{x_1^2 + x_2^2}$

$$\|x\|_2^2 = \sum_{i=1}^{n} x_i^2$$

squared two-norm / Euclidean norm

Assumption: there exist weights $\beta \in \mathbb{R}^d$ such that

$$x_i^T \beta \approx y_i \quad \text{for all these observations}$$

Choose

$$J(\beta) = \frac{1}{n} \sum_{i=1}^{n} (x_i^T \beta - y_i)^2$$

(Ordinary Least Squares)

$$\beta_{OLS} = \underset{\beta}{\operatorname{argmin}}\, J(\beta) = \underset{\beta}{\operatorname{argmin}}\, \frac{1}{n} \| X\beta - y \|_2^2$$

Ordinary Least Squares

Fact: $\|x\|_2^2 = \langle x, x \rangle = x^T x$

$$\Rightarrow \frac{1}{n}\|X\beta - y\|_2^2 = \frac{1}{n}\langle X\beta - y, X\beta - y \rangle$$

Fact: $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$

$$\Rightarrow \frac{1}{n}\langle X\beta, X\beta - y \rangle + \frac{1}{n}\langle -y, X\beta - y \rangle$$

Fact: $\langle x, y \rangle = \langle y, x \rangle$

$$\Rightarrow \frac{1}{n}\langle X\beta - y, X\beta \rangle + \frac{1}{n}\langle X\beta - y, -y \rangle$$

$$\Rightarrow \quad = \frac{1}{n}\langle X\beta, X\beta \rangle + \frac{1}{n}\langle -y, X\beta \rangle + \frac{1}{n}\langle X\beta, -y \rangle + \frac{1}{n}\langle -y, -y \rangle$$

Fact: $\langle cx, z\rangle = c\langle x, z\rangle$

$$\Rightarrow \quad = \frac{1}{n}\langle X\beta, X\beta\rangle - \frac{1}{n}\langle y, X\beta\rangle - \frac{1}{n}\langle X\beta, y\rangle + \frac{1}{n}\langle y, y\rangle$$

$$= \frac{1}{n}\langle X\beta, X\beta\rangle - \frac{2}{n}\langle X\beta, y\rangle + \frac{1}{n}\langle y, y\rangle$$

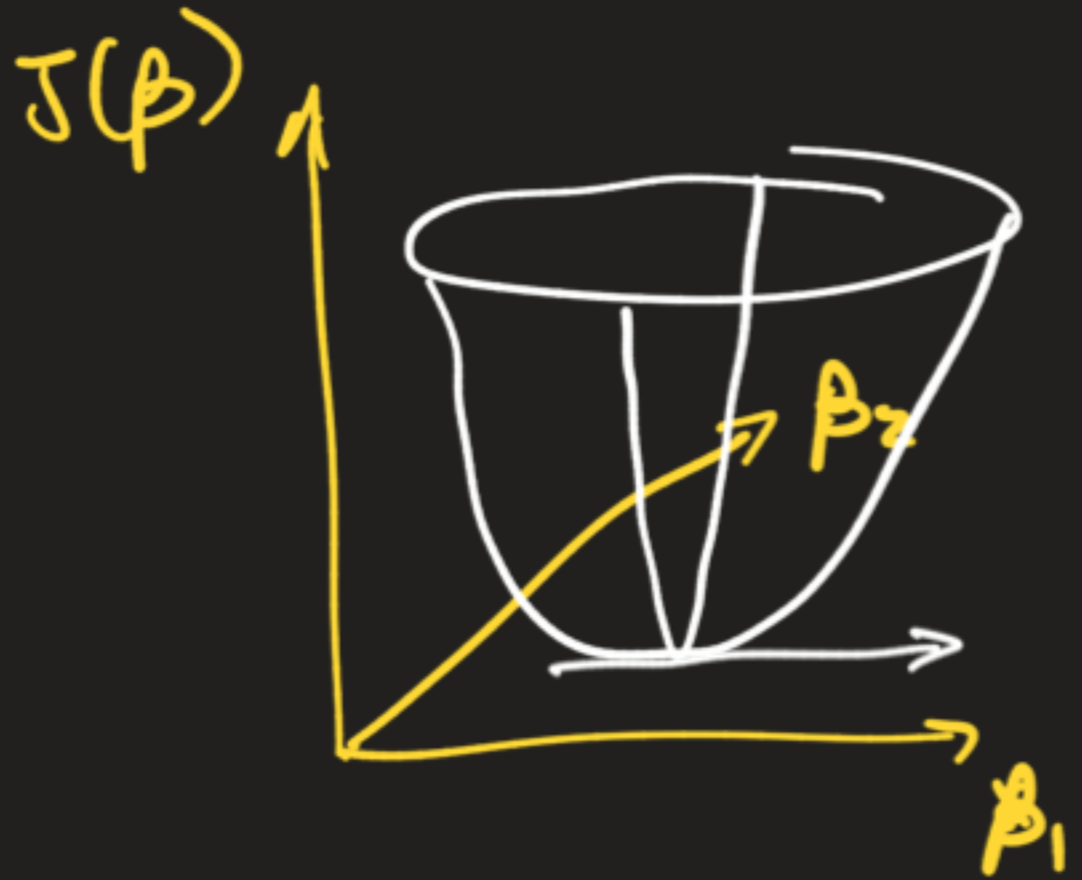
$$\boxed{J(\beta) = \frac{1}{n}\beta^T X^T X\beta - \frac{2}{n}y^T X\beta + \frac{1}{n}y^T y}$$

Recall from calculus: the gradient of a continuously differentiable function is zero at any maximizer or minimizer

$$\Rightarrow \quad \nabla_\beta J(\beta_{OLS}) = 0$$

Recall the gradient of a function $J: \mathbb{R}^d \to \mathbb{R}$ is the vector

$$\nabla_\beta J(\beta) = \begin{bmatrix} \frac{\partial J}{\partial \beta_1} \\ \vdots \\ \frac{\partial J}{\partial \beta_d} \end{bmatrix} \in \mathbb{R}^d$$



Q: what is $\nabla_\beta J$?

<u>Fact</u>: $\nabla_\beta \langle \beta, z \rangle = \nabla_\beta \left(\sum_{i=1}^{d} \beta_i z_i \right) = \nabla_\beta \left(\beta_1 z_1 + \cdots + \beta_d z_d \right)$

$$= \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_d \end{bmatrix} = z$$

$$\boxed{\nabla_\beta \langle \beta, z \rangle = z}$$

<u>Fact</u>: $\nabla_\beta \left[\beta^T M \beta \right] = \nabla_\beta \left(\sum_{i=1}^{d} \sum_{j=1}^{d} M_{ij} \beta_i \beta_j \right)$

$$= \nabla_\beta \left(M_{11} \beta_1^2 + M_{12} \beta_1 \beta_2 + M_{21} \beta_1 \beta_2 + \cdots \right)$$

Note that $\sum_{i=1}^{d}\sum_{j=1}^{d} M_{ij}\beta_i\beta_j =$

$$\underbrace{\underbrace{M_{11}\beta_1^2}_{i=j=1} + \underbrace{\sum_{i=2}^{d} M_{i1}\beta_i\beta_1}_{\substack{i\neq 1\\ j=1}} + \underbrace{\sum_{j=2}^{d} M_{1j}\beta_1\beta_j}_{\substack{j\neq 1\\ i=1}}}_{\text{only these terms depend on } \beta_1} + \underbrace{\sum_{i=2}^{d}\sum_{j=2}^{d} M_{ij}\beta_i\beta_j}_{\substack{i\neq 1\\ j\neq 1}}$$

$\Rightarrow$

$$\frac{\partial[\beta^T M\beta]}{\partial\beta_1} = 2M_{11}\beta_1 + \sum_{i=2}^{d} M_{i1}\beta_i + \sum_{j=2}^{d} M_{1j}\beta_j$$

$$= \sum_{i=1}^{d} M_{i1}\beta_i + \sum_{j=1}^{d} M_{1j}\beta_j = \sum_{i=1}^{d} (M^T)_{1i}\beta_i + (M\beta)_1$$

$$= (M^T\beta)_1 + (M\beta)_1$$

$$= [(M+M^T)\beta]_1$$

similarly,

$$\frac{\partial[\beta^T M\beta]}{\partial\beta_j} = [(M+M^T)\beta]_j$$

for $j=2,\dots,d$

and we find $\nabla_\beta(\beta^T M \beta) = (M + M^T)\beta$

in our case $M = X^T X = M^T$, so

$$\nabla_\beta(\beta^T X^T X \beta) = 2X^T X\beta$$

and

$$\Rightarrow \nabla_\beta J(\beta) = \frac{2}{n} X^T X\beta - \frac{2}{n} X^T y = \frac{2}{n} X^T (X\beta - y)$$

so the fact that $\nabla_\beta J(\beta_{OLS}) = 0$ implies

$$X^T X \beta_{OLS} = X^T y$$

so

$$\beta_{OLS} = (X^T X)^{-1} X^T y$$

Ex: NNLS (nonnegative least squares)

$$\beta^* = \underset{\beta \geq 0}{\operatorname{argmin}} \|X\beta - y\|_2^2$$

This constrained optimization problem does NOT have a closed form solution. We use iterative algorithms (e.g. projected gradient descent):

$\beta_0 \leftarrow$ arbitrary initial point

for $t = 1$ to T:

$$\beta_t \leftarrow \left[\beta_{t-1} - \alpha \nabla_\beta J(\beta_{t-1})\right]_+ = \left[\beta_{t-1} - \alpha X^T(X\beta_{t-1} - y)\right]_+$$

where $[x]_+$ is the operation that sets negative entries to zero