EECS 127/227AT Optimization Models in Engineering
Spring 2020 Lecture 2/18

Today: Connections.
1. Optimization – Probability
2. Principal Components Regression
3. Total Least Square

Ridge Regression:

$$\min ||X\vec{w} - \vec{y}||_2^2 + \lambda^2 ||\vec{w}||_2^2$$

How can we use probabilistic information about our data? How does this connect to optimization models? $(\vec{x_i}, y_i)$ are my data points.

$$y_i = g(\vec{x_i}) + z_i$$

i.i.d.

$$z_i \sim N(0, \sigma_i^2)$$

$$f_{z_i}(z_i) = \frac{e^{-z_i^2/2\sigma_i^2}}{\sqrt{2\pi}\sigma_i}$$

Consider linear model:

$$y_i = \vec{x_i}^T \vec{w} + z_i$$

$\vec{w}$ is our "model": unknown and what we want to learn.

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} = \begin{bmatrix} \vec{x_1}^T \\ \cdot\cdot \\ \cdot\cdot \\ \vec{x_n}^T \end{bmatrix} \vec{w} + \begin{bmatrix} z_1 \\ z_2 \\ \cdot\cdot \\ z_n \end{bmatrix}$$

In a more concise form:

$$\vec{y} = X\vec{w} + \vec{z}$$

Probabilistic Solution::

Maximum likelihood estimator
Find that $\vec{w}$ that makes the observed data most likely.

$$\underset{\vec{w_0}}{\operatorname{argmax}} f_{y_1 y_2 \ldots y_n}(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n | \vec{w} = \vec{w_0})$$

(Maximum Likelihood)

$$= \underset{\vec{w_0}}{\operatorname{argmax}} \Pi_{i=1}^{n} f(Y_i = y_i | \vec{w} = \vec{w_0})$$

<u>Consider:</u>

$$f(Y_i = y_i|\vec{w} = \vec{w_0}) = f(\vec{x_i}^T \vec{w_0} + z_i = y_i|\vec{w} = \vec{w_0})$$

(Because all of my z_i's are independent)

<u>Consider:</u>

$$f(Y_i = y|\vec{w} = \vec{w_0}) = f(\vec{x_i}^T \vec{w_0} + z_i = y_i|\vec{w} = \vec{w_0})$$

$$= f(z_i = y_i - \vec{x_i}^T \vec{w_0}|\vec{w} = \vec{w_0}) = \frac{e^{-(y_i - \vec{x_i}\vec{w_0})^2/\sigma_i^2}}{\sqrt{2\pi}\sigma_i}$$

Then we want to find

$$\underset{\vec{w_0}}{\operatorname{argmax}} \Pi_{i=1}^{n} \frac{e^{-(y_i - \vec{x_i}^T \vec{w_0})^2/2\sigma_i^2}}{\sqrt{2\pi}\sigma_i}$$

$$= \underset{\vec{w_0}}{\operatorname{argmax}} \frac{1}{(\sqrt{2\pi})^n} \frac{1}{\Pi_{i=1}^{n}\sigma_i} \exp\left(\Sigma_{i=1}^{n} - (y_i - \vec{x_i}^T \vec{w_0})^2/2\sigma_i^2\right)$$

$$= \underset{\vec{w_0}}{\operatorname{argmax}} \Sigma_{i=1}^{n}(y_i - \vec{x_i}^T \vec{w_0})^2/2\sigma_i^2$$

$$= \underset{\vec{w_0}}{\operatorname{argmax}} ||S(\vec{y} - X\vec{w_0})||^2$$

(weighted least square) Where

$$S^2 = \begin{bmatrix} \frac{1}{2\sigma_1^2} & \dots & 0 & & \\ & \frac{1}{2\sigma_2^2} \dots & 0 & & \\ & & & .. & \\ & & & 0 & \frac{1}{2\sigma_n^2} \end{bmatrix}$$

$$S = \begin{bmatrix} \frac{1}{\sqrt{2}\sigma_1^2} & \dots & 0 & \\ & \frac{1}{\sqrt{2}\sigma_2^2} \dots & 0 & \\ & & & .. \\ 0 & & 0 & \frac{1}{\sqrt{2}\sigma_n^2} \end{bmatrix}$$

What if we had a prior on $\vec{w}$? "side information"

<u>MAP</u>: Maximum a posterior

$$y_i = \vec{x_i}^T \vec{w} + z_i$$

$$z_i \sim N(0, \sigma_i^2)$$

$$w_i \sim N(\mu_i, \delta_i^2)$$

"prior" on $\vec{w}$

$$\vec{w} \sim N(\vec{\mu}, \Sigma_w)$$

Where

$$\vec{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ .. \\ \mu_n \end{bmatrix}$$

$$\Sigma_w = \begin{bmatrix} \delta_1^2 & & & \\ & \delta_2 & & 0 \\ .. & & & \\ & & 0 & \delta_n^2 \end{bmatrix}$$

$$\underset{\vec{w}}{\text{argmax}} \;\; f(\vec{w}|Y_i = y_1, y_2 = y_2, ..., Y_n = y_n) \quad (*)$$

What is the most likely $\vec{w}$ given the data?

$$f(\vec{w}|Y_1 = y_1, ..., Y_n = y_n) = \frac{f(Y_1 = y_1, ..., Y_n = y_n|\vec{w})f(\vec{w})}{f(Y_1 = y_1, ..., Y_n = y_n)}$$

Note this is from Bayes Rule, and the denominator does not depend on $\vec{w}$

$$(*) \text{ MAP} = \underset{\vec{w}}{\text{argmax}}\, f(Y_1 = y_1, ..., Y_n = y_n|\vec{w}) * f(\vec{w})$$

$$= \underset{\vec{w}}{\text{argmax}}\, f(\vec{Y} = \vec{y}|\vec{w}) * f(\vec{w})$$

$$= \underset{\vec{w}}{\text{argmax}} \; [\Pi_{i=1}^n \; f(\vec{Y} = \vec{y}|\vec{w})] * f(\vec{w})$$

$$= \underset{\vec{w}}{\text{argmax}} \; [\Pi_{i=1}^n \frac{\exp{(-\frac{(\vec{x_i}^T\vec{w} - y_i)^2}{2\sigma_i^2})}}{\sqrt{2\pi}} * \sigma_i] * \frac{e^{-(\vec{w} - \vec{\mu}\Sigma_w^{-1}(\vec{(w)} - \vec{\mu})}}{(\sqrt{2\pi})^n(\Pi\delta_i)}$$

$$= \underset{\vec{w}}{\text{argmax}} \;\; \exp \Sigma_{i=1}^n - \frac{(\vec{x_i}^T \underline{w} y_i)^2}{2\sigma_i^2} + -(\vec{w} - \vec{\mu})^T \Sigma_w^{-1}(\vec{w} - \vec{\mu})$$

$$= \text{argmin}\, ||S(X\vec{w} - \vec{y}||_2^2 + ||\sqrt{\Sigma_w^{-1}}(\vec{w} - \vec{\mu})||_2^2$$

What happens if δ_i is large? Choose penalty for deviation from the mean.

Principal Components Regression:

$$\min ||X\vec{w} - \vec{y}||_2^2$$

Where $X \in R^{mxn}$, X is full column rank.
$X = Y\Sigma V^T$ LS:

$$\vec{\hat{w}} = (X^T X)^{-1} X^T \vec{y}$$

$$= ((U\Sigma V^T)^T (U\Sigma V^T))^{-1} (U\Sigma V^T)^{-1} \vec{y}$$

...(Usual Math)...

$$= V \begin{bmatrix} \frac{1}{\sigma_1} & 0 & 0 & 0 \\ & & ... & 0 \\ & 0 & ...\frac{1}{\sigma_n} & 0 \end{bmatrix} U^T \vec{y}$$

For PCA: only consider top k principal components instead of all of X.

Ridge Regression as soft PCA

$$\underset{\vec{w}}{\operatorname{argmin}} ||X\vec{w} - \vec{y}||_2^2 + \lambda ||\vec{w}||_2^2$$

$$= \underset{\vec{w}=V\vec{z}}{\operatorname{argmin}} ||XV\vec{z} - \vec{y}||_2^2 + \lambda ||\vec{z}||_2^2$$

(Ridge)

$$XV = A$$

$$\vec{z_{ridge}} = ((XV)^T (XV) + \lambda I)^{-1} (XV)^T \vec{y}$$

$$= (\Sigma^T \Sigma + \lambda I)^{-1} \Sigma^T U^T \vec{y}$$

$$= (\begin{bmatrix} \sigma_1^2 + \lambda & 0 & 0 \\ & & ... \\ & 0 & ...\sigma_n^2 + \lambda \end{bmatrix})^{-1} \begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ & & ... & 0 \\ & 0 & ...\sigma_n & 0 \end{bmatrix} U^T \vec{y}$$