- EECS 127 Lecture 9

Feb 18, 2020

Today : Connections.

- Optimization $\leftrightarrow$ Probability.
- Principal Components Regression.
- TLS..

Hyperparameter

Ridge regression: minimize: $\|X\vec{w} - \vec{y}\|_2^2 + \lambda \|\vec{w}\|_2^2$

How can we use probablistic information about our data?
How does this connect to optimization models?

$(\vec{x}_i, y_i)$ are my data points.

$y_i = g(\vec{x}_i) + z_i$ iid Independent Identically distributed.

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

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

Consider linear model:

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

$\vec{w}$ is "our model"
What we want to learn. (unknown)

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \vec{x}_1^{\,T} \\ \vdots \\ \vec{x}_n^{\,T} \end{bmatrix} \vec{\omega} + \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{bmatrix} \qquad \vec{y} = X\vec{\omega} + \vec{Z}$$

Probablistic solution: Maximum likilhood estimator.

Find that $\vec{\omega}$ that makes the observed data most likely.

$$\underset{\vec{\omega}_0}{\operatorname{argmax}} \; f_{Y_1 Y_2 \cdots Y_n}\left(Y_1 = y_1, Y_2 = y_2 \cdots, Y_n = y_n \,\middle|\, \vec{\omega} = \vec{\omega}_0\right) \qquad \text{(Maximum Liklihood)}$$

$$= \underset{\vec{\omega}_0}{\operatorname{argmax}} \; \prod_{i=1}^{n} \underbrace{f\left(Y_i = y_i \,\middle|\, \vec{\omega} = \vec{\omega}_0\right)}$$

(Because all of my z_i's are independent.)

Consider: $$f\left(Y_i = y_i \,\middle|\, \vec{\omega} = \vec{\omega}_0\right) = f\left(\vec{x}_i^{\,T}\vec{\omega}_0 + z_i = y_i \,\middle|\, \vec{\omega} = \vec{\omega}_0\right)$$

$$= f\left(z_i = y_i - \vec{x}_i^{\,T}\vec{\omega}_0 \,\middle|\, \vec{\omega} = \vec{\omega}_0\right) = \frac{e^{-(y_i - \vec{x}_i^{\,T}\vec{\omega}_0)^2 / 2\sigma_i^2}}{\sqrt{2\pi}\,\sigma_i} \qquad (*)$$

$$\underset{\vec{w}_0}{\operatorname{argmax}} \prod_{i=1}^{n} \frac{e^{-(y_i - \vec{x}_i^T \vec{w}_0)^2 / 2\sigma_i^2}}{\sqrt{2\pi}\,\sigma_i}$$

$$= \underset{\vec{w}_0}{\operatorname{argmax}} \frac{1}{(\sqrt{2\pi})^n} \frac{1}{\prod_{i=1}^{n} \sigma_i} \exp\left\{ \sum_{i=1}^{n} -(y_i - \vec{x}_i^T \vec{w}_0)^2 / 2\sigma_i^2 \right\}$$

$$= \underset{\vec{w}_0}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \vec{x}_i^T \vec{w}_0)^2 / 2\sigma_i^2$$

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

$$= \underset{\vec{w}_0}{\operatorname{argmin}} \underbrace{\| S(\vec{y} - X\vec{w}_0) \|^2}_{\text{Weighted Least Squares.}}$$

$$S = \begin{bmatrix} \frac{1}{\sqrt{2}\sigma_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sqrt{2}\sigma_n} \end{bmatrix}$$

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

MAP: Maximum a posteriori

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

$$w_i \sim N(\mu_i, \rho_i^2) \quad \text{"Prior" on } \vec{w}$$

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

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

$$\vec{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} ; \quad \Sigma_w = \begin{bmatrix} \rho_1^2 & & & 0 \\ & \rho_2^2 & & \\ & & \ddots & \\ 0 & & & \rho_n^2 \end{bmatrix}$$

$$\underset{\vec{w}}{\operatorname{argmax}}\, f(\vec{w} \mid Y_1 = y_1, Y_2 = y_2 \ldots Y_n = y_n) \quad (*)$$

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

$$f(\vec{w} \mid Y_1 = y_1 \ldots Y_n = y_n) = \frac{f(Y_1 = y_1, Y_2 = y_2 \ldots Y_n = y_n \mid \vec{w}) \cdot f(\vec{w})}{f(Y_1 = y_1, Y_2 = y_2 \ldots, Y_n = y_n)} \quad \text{(Bayes Rule)}$$

does not depend on $\vec{w}$.

$$(*)\ \text{MAP} = \underset{\vec{w}}{\operatorname{argmax}}\, f(Y_1 = y_1, \ldots Y_n = y_n \mid \vec{w}) \cdot f(\vec{w})$$

$$\vec{Y} = \vec{y}$$

$$= \underset{\vec{w}}{\operatorname{argmax}}\, f(\vec{Y} = \vec{y} \mid \vec{w})\, f(\vec{w})$$

$$= \underset{\vec{w}}{\operatorname{argmax}} \left(\prod_{i=1}^{n} f(Y_i = y_i \mid \vec{w}) \right) f(\vec{w}).$$

$$= \underset{\vec{w}}{\operatorname{argmax}} \left(\prod_{i=1}^{n} \frac{\exp\left(-\frac{(\vec{x}_i^{T}\vec{w}-y_i)^2}{2\sigma_i^2}\right)}{\sqrt{2\pi}\cdot\sigma_i} \right) \cdot \frac{e^{-(\vec{w}-\vec{\mu})^T \Sigma_w^{-1}(\vec{w}-\vec{\mu})}}{(\sqrt{2\pi})^n \left(\prod \varsigma_i\right)}$$

$$= \underset{\vec{w}}{\operatorname{argmax}} \exp\left\{ \sum_{i=1}^{n} -\frac{(\vec{x}_i^{T}\vec{w}-y_i)^2}{2\sigma_i^2} + -(\vec{w}-\vec{\mu})^T \Sigma_w^{-1}(\vec{w}-\vec{\mu}) \right\}$$

$$= \operatorname{argmin} \|S(X\vec{w}-\vec{y})\|_2^2 + \|\underbrace{\sqrt{\Sigma_w^{-1}}}_{\text{like the } \lambda \text{ terms.}}(\vec{w}-\vec{\mu})\|_2^2$$

What happens if ς_i is large?

choose less penalty for deviation from the mean.

Principal Components Regression.

$X \in \mathbb{R}^{m \times n}$.

X is full column rank.

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

$X = U \Sigma V^T$.

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

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

$\vdots$ Do usual math.

$$= V \left[\begin{array}{ccc|c} \frac{1}{\sigma_1} & & 0 & \\ & \ddots & & 0 \\ 0 & & \frac{1}{\sigma_n} & \end{array} \right] U^T \vec{y}.$$

For PCR: Only consider top k principal components instead of all of X.

"Ridge regression as soft PCA"

$\vec{w}$ in the V basis.

$$\vec{w} = V \cdot \vec{z}$$

$$X = U \Sigma V^T$$

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

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

$$= \underset{\vec{z}}{\operatorname{argmin}} \; \| \underbrace{XV}_{A} \vec{z} - \vec{y} \|_2^2 + \lambda \| \vec{z} \|_2^2 \quad \text{(Ridge).}$$

$$(A^T A + \lambda I) A^T \vec{b}$$

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

$$= (V^T X^T X V + \lambda I)^{-1} (XV)^T \cdot \vec{y}$$

$$\cancel{= (V^T V^T \Sigma U^T U}$$

$$= (V^T (U \Sigma V^T)^T (U \Sigma V^T) V + \lambda I)^{-1} (XV)^T \cdot \vec{y}$$ } HW: do this cancellation.

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

$$= \begin{pmatrix} \sigma_1^2 + \lambda & & & \\ & \sigma_2^2 + \lambda & & \\ & & \ddots & \\ & & & \sigma_n^2 + \lambda \end{pmatrix}^{-1} \left[\begin{array}{cccc|c} \sigma_1 & & & & \\ & \sigma_2 & & & \\ & & \ddots & & 0 \\ & & & \sigma_n & \end{array}\right] U^T \vec{y}$$