Today: connections

- Maximum likelihood estimation
- Maximum a-posteriori
- Tikhamov regression

Recall: Ridge Regression

$$\min_{\vec{x}} \|A\vec{x}-\vec{b}\|^2 + \underbrace{\lambda^2\|\vec{x}\|^2}_{\text{regularizer}} \longrightarrow \boxed{\vec{x}^* = (A^TA+\lambda I)^{-1}A^T\vec{b}}$$

→ Note! the solution to this $\neq$ solution to the least squares problem

↳ Multiple interpretations of this optimization problem

(1) Wanted to be robust to perturbations & reduce sensitivity towards them

↳ wanted to make sure $\sigma_i\{A\}$ not too large ⇒ shifted them away from 1

→ w/o this ridge coefficient/lambda term, our predicted coefficients were really large

↳ said we knew that $\vec{x}$ wasn't very large

(2) Ghost data: Add measurements to least squares setup saying how confident we were about closeness to $\vec{0}$

Tikhonov Regularization

↳ generalization of ridge

$A\vec{x} = \vec{b}$

$$\begin{bmatrix} A \\ \lambda I \end{bmatrix} \begin{bmatrix} \vec{x} \end{bmatrix} = \begin{bmatrix} \vec{b} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \text{ridge} \longrightarrow \quad \begin{bmatrix} W_1 A \\ W_2 \end{bmatrix} \begin{bmatrix} \vec{x} \end{bmatrix} = \begin{bmatrix} \vec{b} \\ \vec{x}_0 \end{bmatrix}$$

idea: say we know more about the structure of the matrices; can we weight them by some values to tell us more about our data?

Tikhonov regularization

$$\boxed{\min_{\vec{x}} \|W_1(A\vec{x}-\vec{b})\|_2^2 + \|W_2(\vec{x}-\vec{x}_0)\|_2^2}$$

Probalistic Perspective

- to figure out what side information to incorporate

$$y_i = g(x_i) + z_i$$

↖ data points

- say we have a linear model:

$$g(\vec{x}_i) = \vec{x}_i^T \vec{\omega}$$

↳ $\vec{\omega}$ is our model

- say noise is drawn from a Normal distribution: $\mathcal{Z}_i \sim N(0, \sigma_i^2)$

Recall: The density fcn for a normal distribution:

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

→ assuming these are all iid (independent, ~~identically~~ distributed)

↓

can rewrite this as

$$y_i = g(\vec{x}_i^T\vec{\omega}) + z_i$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} — \vec{x}_1^T — \\ \vdots \\ — \vec{x}_n^T — \end{bmatrix} \begin{bmatrix} | \\ \vec{\omega} \\ | \end{bmatrix} + \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix}$$

$\vec{y} \approx X\vec{w}$

↳ Note: we can solve this with least squares but that doesn't tell us everything about our measurements; don't take the distribution of our noise into account

solution!
MLE

Maximum Likelihood estimation: $\vec{w}$ that makes observed

$$\underset{\vec{w}_0}{\operatorname{argmax}}\ f(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n \mid \underbrace{\vec{w} = \vec{w}_0}_{\text{we're assuming we know nothing about } \vec{w} \text{ (no priori)}})$$

↳ ask us: what is the unknown that makes our data most likely?

→ trying to maximize the density in the case of continuous RV's

• since z_i's are iid, we can write y_i's as a product (because the only thing they're contingent on is the $\vec{w}$ but we're conditioning that out):

$$\underset{\vec{w}_0}{\operatorname{argmax}}\ f(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n \mid \vec{w} = \vec{w}_0)$$

$$= \underset{\vec{w}_0}{\operatorname{argmax}} \prod_{i=1}^{n} \boxed{f(Y_i = y_i \mid \vec{w} = \vec{w}_0)}$$

↳ want to understand this conditional density

Ayan's attempt

$F(Y_i = y_i \mid \vec{w} = \vec{w}_0)$

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

↳ use Baye's rule? idk

Ranade

$$F(Y_i = y_i \mid \vec{w} = \vec{w}_0) = F(\vec{x}_i^{\,T}\vec{w}_0 + z_i = y_i \mid \vec{w} = \vec{w}_0)$$

$$= F(z_i = y_i - \vec{x}_i^{\,T}\vec{w}_0 \mid \vec{w} = \vec{w}_0)$$

$$= \frac{e^{-\frac{(y_i - \vec{x}_i^{\,T}\vec{w}_0)^2}{2\sigma_i^2}}}{\sqrt{2\pi}\cdot\sigma_i}$$

can now rewrite this as

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

↳ since denominator not contingent on $\vec{\omega}_0$, can rewrite as:

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

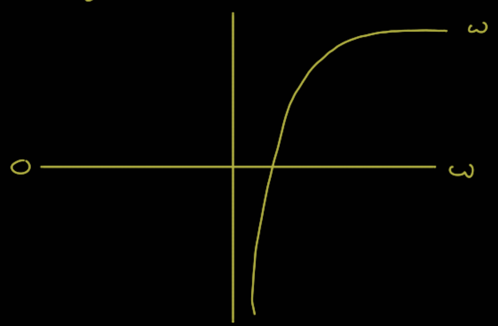
$$\prod_{i=1}^{n} e^{x_i} = e^{x_1}e^{x_2}\cdots e^{x_n} = e^{\sum_i^n x_i} = \exp\left\{\sum_{i=1}^{n} x_i\right\}$$

Recall:

- $\underset{\vec{x}}{\operatorname{argmax}} \|A\vec{x} - \vec{b}\|_2 = \underset{\vec{x}}{\operatorname{argmax}} \|A\vec{x} - \vec{b}\|_2^2$

⋮

- $\underset{\omega}{\operatorname{argmax}}\ F(\omega) = \operatorname{argmax} \log F(\omega)$



because it's monotonically increasing. This is the case for any increasing fcn.

IMPORTANT

$$\max_{\vec{x}} \|A\vec{x} - \vec{b}\|_2 \neq \max_{\vec{x}} \|A\vec{x} - \vec{b}\|_2^2$$

lets us rewrite as

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

taking logs & dropping constants

$$= \underset{\vec{\omega}_0}{\operatorname{argmax}} \left\{-\sum_{i=1}^{n} \frac{(\vec{y}_i - \vec{x}_i^T\vec{\omega}_0)^2}{2\sigma_i^2}\right\}$$

$$= \underset{\vec{\omega}_0}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{(\vec{y}_i - \vec{x}_i^T\vec{\omega}_0)^2}{\boxed{2\sigma_i^2}}$$

Recall: our least-squares cost is $\sum_{i=1}^{n}(x_i^T\vec{\omega} - y_i)^2$

this is the same as our least squares cost except it's $\boxed{\text{weighted}}$ by some value

$$= \underset{\vec{\omega}_0}{\operatorname{argmin}} \|S(X\vec{\omega}_0 - \vec{y})\|_2^2$$

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

MAP (Maximum a Posteriori)

↳ used if we have a prior on $\vec{\omega}$

$$\vec{y}_i = \vec{x}_i^T \vec{\omega} + z_i \qquad z_i \sim N(0, \sigma_i^2)$$

$$\omega_i \sim N(\mu_i, \rho_i^2)$$

$$\vec{\omega} \sim N(\vec{\mu}, \Sigma_\omega)$$

$$\underbrace{\Sigma_{\vec{\omega}}}_{\text{covariance matrix}} = \begin{bmatrix} \rho_1^2 & & & 0 \\ & \rho_2^2 & & \\ & & \ddots & \\ 0 & & & \rho_n^2 \end{bmatrix}$$

↳ variances along diagonal

→ 0 covariance

(w_i not correlated w/ w_j if $i \neq j$)

Recall from IGA, imaging:

↳ if we have

What is most likely data given $y_1, y_2, \dots, y_n$?

$$\arg\max_{\vec{\omega}} f(\vec{\omega} \mid Y = \vec{y})$$

↖ in MLE, these were swapped

→ using Baye's rule, rewrite as:

$$f(\vec{\omega} \mid Y = \vec{y}) = \frac{f(Y = \vec{y} \mid \vec{\omega}) f(\vec{\omega})}{\boxed{f(\vec{y})}}$$

↳ this is a constant; your data is what it is (can ignore it as a constant in our argmax)

$$= \arg\max_{\vec{\omega}} f(Y = \vec{y} \mid \vec{\omega}) f(\vec{\omega})$$

$$= \boxed{\arg\max_{\vec{\omega}} \prod_{i=1}^{n} f(Y = y_i \mid \vec{\omega})} \underbrace{f(\vec{\omega})}_{\text{just the multivariate Gaussian density}}$$

already computed this previously

Multivariate normal

$$w \sim \frac{e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\,\sigma^2}$$

$$= \arg\max_{\vec{\omega}} \prod_{i=1}^{n} \frac{\exp\left\{ \frac{-(y_i - \vec{x}_i^T \vec{\omega})^2}{2\sigma_i^2} \right\}}{\sqrt{2\pi}\,\sigma_i} \cdot \frac{\exp\left\{ -(\vec{\omega} - \vec{\mu})^T \Sigma_\omega^{-1} (\vec{\omega} - \vec{\mu}) \right\}}{(\sqrt{2\pi})^n \left(\prod_{i=1}^{n} \rho_i\right)}$$

neglible constants

$$= \arg\max_{\vec{\omega}} \exp\left\{ -\sum_{i=1}^{n} \frac{(y_i - \vec{x}_i^T \vec{\omega})^2}{2\sigma_i^2} + -(\vec{\omega} - \vec{\mu})^T \Sigma_\omega^{-1} (\vec{\omega} - \vec{\mu}) \right.$$

$$= \arg\min_{\vec{\omega}} \sum_{i=1}^{n} \frac{(y_i - \vec{x}_i^T \vec{\omega})^2}{2\sigma_i^2} + (\vec{\omega} - \vec{\mu})^T \Sigma_\omega^{-1} (\vec{\omega} - \vec{\mu})$$

Because Σ_ω is symmetric can write as

$$\Sigma_\omega^{-1} = \sqrt{\Sigma_\omega^{-1}} \sqrt{\Sigma_\omega^{-1}}$$

$$= \arg\min_{\vec{\omega}} \| S(X\vec{\omega} - \vec{y}) \|_2^2 + \left\| \sqrt{\Sigma_\omega^{-1}} (\vec{\omega} - \vec{\mu}) \right\|_2^2$$

↳ if $\vec{\mu} = \vec{0}$ we'd get ridge regression