Generative Models (Just-so-stories)

Suppose we know $y = w_\star^T x + \text{noise}$

$w_\star$ "true" model, noise Gaussian w/ mean zero and variance ν^2.

In vector form $\vec{y} = \bar{X} w_\star + \vec{n}$

(this is never really true, but is instructive)

~~What is~~

PREDICTION ERROR: Let $y_i^{true} = w_\star^T x_i$

For a learned model $\hat{w}$, $y_i^{pred} = \hat{w}^T x_i$

prediction error: $\frac{1}{n} \sum_{i=1}^{n} (y_i^{pred} - y_i^{true})^2 = \frac{1}{n} \| \bar{X}(\hat{w} - \vec{w}_\star)\|^2$

OLS: $\frac{1}{n} \sum_{i=1}^{n} (y_i^{pred} - y_i^{true})^2 = \frac{1}{n} \| \bar{X} (\vec{w}_{OLS} - \vec{w}_\star)\|^2$

$\vec{w}_{OLS} = (\bar{X}^T \bar{X})^{-1} \bar{X}^T \vec{y}$ (assume $d \leq n$ and $\bar{X}$ full rank)

$$\begin{aligned}
\bar{X}(\vec{w}_{OLS} - \vec{w}_\star) &= \bar{X}\left((\bar{X}^T\bar{X})^{-1}\bar{X}^T\vec{y} - w_\star\right) \\
&= \bar{X}\left((\bar{X}^T\bar{X})^{-1}\bar{X}^T(\bar{X}\vec{w}_\star + \vec{n}) - w_\star\right) \\
&= \bar{X}(\bar{X}^T\bar{X})^{-1}\bar{X}^T\vec{n}
\end{aligned}$$

$$\bar{X} = \bar{U}\bar{S}\bar{V}^T$$

$$\bar{X}(\bar{X}^T\bar{X})^{-1}X^T = (\bar{U}\bar{S}\bar{V}^T)(\bar{V}\bar{S}^T\bar{S}\bar{V}^T)^{-1}(\bar{V}\bar{S}^T\bar{U}^T)$$

$$= (\bar{U}\bar{S}\bar{V}^T)(\bar{V}(\bar{S}^T\bar{S})^{-1}\bar{V}^T)(\bar{V}\bar{S}^T\bar{U}^T)$$

$$= \bar{U}\bar{S}(\bar{S}^T\bar{S})^{-1}S^T\bar{U}^T$$

$$= \bar{U}\begin{bmatrix}\bar{I}_d & \bar{0}\\ \bar{0} & \bar{0}\end{bmatrix}U^T$$

$$\mathbb{E}\left[\frac{1}{n}\|\bar{X}(\vec{w}_{OLS} - \vec{w}_*)\|^2\right] = \frac{1}{n}\mathbb{E}\left[\|\bar{U}\begin{bmatrix}\bar{I}_d & \bar{0}\\ \bar{0} & \bar{0}\end{bmatrix}\bar{U}^T\vec{n}\|^2\right]$$

$$= \frac{1}{n}\mathbb{E}\left[\vec{n}^T\bar{U}\begin{bmatrix}\bar{I}_d & 0\\ \bar{0} & \bar{0}\end{bmatrix}\bar{U}^T\vec{n}\right]$$

$$= \frac{1}{n}\mathbb{E}\,\mathrm{Trace}\left[\begin{bmatrix}\bar{I}_d & \bar{0}\\ \bar{0} & \bar{0}\end{bmatrix}\bar{U}^T\vec{n}\vec{n}^T\bar{U}\right]$$

$$= \frac{1}{n}\,\mathrm{Trace}\left[\begin{bmatrix}\bar{I}_d & \bar{0}\\ \bar{0} & \bar{0}\end{bmatrix}\bar{U}^T\mathbb{E}[\vec{n}\vec{n}^T]\bar{U}\right]$$

$$= \frac{1}{n}\,\mathrm{Trace}\left[\begin{bmatrix}\bar{I}_d & \bar{0}\\ \bar{0} & \bar{0}\end{bmatrix}U^T(\nu^2 I_n)\bar{U}\right]$$

$$= \nu^2\frac{d}{n}$$

PINV: assume $w_{\star} \in \text{range}(\bar{X})$

$$\vec{w}_{PINV} = \bar{X}^{\dagger} \vec{y} \quad , \quad \text{rank}(\bar{X}) = r$$

$$\bar{X}(\vec{w}_{PINV} - \vec{w}_{\star}) = \bar{X}\left[(\bar{X}^T \bar{X}) \bar{X}^T (\bar{X} \bar{w}_{\star} + \vec{n}) - w_{\star} \right]$$

$$= \bar{X}(\bar{X}^T \bar{X}) \bar{X}^T n$$

,

BUT $\mathbb{E}\left[\frac{1}{n} \| X(\vec{w}_{OLS} - w_{\star}) \|^2 \right] = \frac{1}{n} \mathrm{Tr}\left[\begin{bmatrix} \bar{I}_r & \bar{0} \\ \bar{0} & \delta \end{bmatrix} \bar{U}^T (\nu^2 I) \bar{U} \right]$

$$= \frac{\nu^2 r}{n}$$

RIDGE attempts to approximate PINV by choice of γ

Random vectors

$$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

$p(\vec{x})$ is probability density for $\vec{x}$.

$$p(\vec{x}) = \frac{\partial^d}{\partial x_1 \cdots \partial x_d} \Pr[x_1 \le a_1, x_2 \le a_2, \ldots, x_d \le a_d]$$

$$\int p(\vec{x}) = 1$$

$$p(x_1) = \int p(x)\, dx_2\, dx_3 \cdots dx_d$$

<u>Mean:</u> $\mathbb{E}[\vec{x}] = \vec{\mu}$

<u>Covariance:</u> $\bar{\Sigma} = \mathbb{E}[(\vec{x} - \vec{\mu})(\vec{x} - \vec{\mu})^T]$

$$= \mathbb{E}[\vec{x}\vec{x}^T] - \vec{\mu}\vec{\mu}^T$$

Σ_{ii} = variance of x_i

Σ_{ij} = covariance of x_i and x_j

If $\vec{x}$ and $\vec{z}$ are random vectors and

$\vec{u} = \bar{A}\vec{x} + \bar{B}\vec{z} + \vec{c}$, then

$$\mu_{\vec{u}} = \bar{A}\vec{\mu}_x + \bar{B}\vec{\mu}_z + \vec{c}$$

$$\Sigma_{\vec{u}} = \bar{A}\bar{\Sigma}_x\bar{A}^T + \bar{A}\bar{\Sigma}_{xz}\bar{B}^T + \bar{B}\bar{\Sigma}_{zx}\bar{A}^T + \bar{B}\bar{\Sigma}_z\bar{B}^T$$

If $z = \vec{v}^T \vec{x}$ for non random $\vec{v}$ then

$$\bar{\bar{\Sigma}}_z = \sigma_z^2 = \vec{v}^T \bar{\bar{\Sigma}}_x \vec{v} \geq 0$$

$\bar{\bar{\Sigma}}_x$ is positive semidefinite (symmetric, all eigenvalues non-negative)

Least-squares revisited

Suppose $(\vec{x}, y)$ has mean $(0,0)$

Consider $\mathbb{E}[(\vec{w}^T \vec{x} - y)^2] = \text{var}(\vec{w}^T \vec{x} - y)$

$$= \vec{w}^T \Sigma_x w - 2 w^T \Sigma_{xy} + \sigma_y^2$$

Then the minimizer of expected squared loss is

$$\vec{w}_{LS} = \Sigma_x^{-1} \Sigma_{xy} \qquad \left(\text{compare to } (\bar{X}^T \bar{X})^{-1} (\bar{X}^T \vec{y})\right)$$

$$\min_{\vec{w}} \mathbb{E}[(\vec{w}^T \vec{x} - y)^2] = \sigma_y^2 - \vec{\Sigma}_{yx} \overline{\Sigma}_x^{-1} \vec{\Sigma}_{xy}$$

($\vec{w}_{LS}^T \vec{x} - y$ is a random variable w/ mean 0 and variance $\sigma_y^2 - \overline{\Sigma}_{yx} \overline{\Sigma}_x^{-1} \overline{\Sigma}_{xy}$)

Multivariate Gaussian: Everyone's favorite

(i) only need to know first and second moments

(ii) CLT implies most data resembles Gaussians when n is large

$$p(\vec{x}) = \frac{1}{|2\pi \Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1}(\vec{x}-\vec{\mu})\right)$$

$$= \mathcal{N}(\vec{\mu}, \Sigma)$$

$$\int p(\vec{x})\, dx_2, \ldots dx_d = \mathcal{N}(\mu_1, \sigma_1^2)$$

Diagonal Case:

$$p(\vec{x}) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2} \frac{(x_i-\mu_i)^2}{\sigma_i^2}\right)$$

For Gaussians, covariance $= 0 \Rightarrow$ independence.

Isocontours: $p(\vec{x}) = p_0 \iff (\vec{x}-\vec{\mu})^T \Sigma^{-1}(\vec{x}-\vec{\mu}) = c_0$

Level sets of probability are ellipses.

Center is at $\vec{\mu}_x$, Σ^{-1} determine the axes.

$$\Sigma = \sum_{i=1}^{d} \lambda_i \vec{v}_i \vec{v}_i^T ,$$

$\vec{v}_i$ determine major axes of ellipse

λ_i determine lengths

$$\underset{\vec{w}}{\text{minimize}} \quad \vec{w}^T \overline{Q} \vec{w} - \cancel{2} \vec{p}^T \vec{w} + r$$

complete the square

$$(\vec{w} - \overline{Q}^{-1} \vec{p})^T \overline{Q} (\vec{w} - \overline{Q}^{-1} \vec{p})$$
$$\vec{p}^T \overline{Q}^{-1} \vec{p} + r$$

$$\vec{w} = \overline{Q}^{-1} p$$

$$\text{cost} = r - \vec{p}^T Q^{-1} p$$