Criteria based design

Tuesday, January 31, 2023 11:15 PM

- We will look at designs based on a specific criteria : minimizing variance in model parameters

- specifically, we look at couple of tools related to the previous lecture on information theory.

Maximum entropy

- Suppose we have a model whose parameters are β and we specified a prior $P(\beta)$

  we can derive our posterior to be

$$p(\beta \mid y, D) = \frac{P(y \mid \beta, D)\, P(\beta)}{P(y \mid D)}$$

- Observe that we have a prior that is a simple known distribution and are intersted in modeling the posterior after observing some data D that predicts distribution on outputs y

- From the previous lecture, we can define a score function using the KL divergence : posterior relative to the prior

$$\text{score} = \int P(\beta \mid y, D) \log\left(\frac{P(\beta \mid y, D)}{P(\beta)}\right) d\beta\, dy$$

- if we are interested collecting data such that your prior and posterior look similar such that you can use one as a proxy for other, we need to minimize the score.

$$\text{score} = \int P(\beta \mid y, D) \log(P(\beta \mid y, D) - \underbrace{\int P(\beta \mid y, D) \log P(\beta)}_{\text{constant}})$$

$$= \min\{-H(P(\beta \mid y, D))\}$$

$$= \max\{H(P(\beta \mid y, D))\}$$

(Thus fitting your posterior to the prior is maximizing the entropy of obtaining parameters β from posterior)

example: one dimensional linear regression

model : $y = A^T x + \varepsilon \sim N(A^T x, \sigma^2)$

$$P(y \mid x, \beta) = N(A^T x, \sigma^2) \qquad \beta = [A, \sigma^2]$$

posterior

$$p(\beta \mid y, x) = P(y \mid x, \beta)\, P(\beta)$$

$$\propto \prod_{i=1}^{n} p(y_i \mid x_i, \beta)$$

$$= \prod_{i=1}^{n} N(A^T x_i, \sigma^2)$$

$$\propto \exp\Big(\underbrace{\frac{-1}{2\sigma^2} \sum_{i=1}^{n} (y_i - A^T x_i)}_{T}\Big)$$

$$= \exp(-T)$$

$$\text{score} = -\mathbb{E}[\log P(\beta \mid y, x)]$$

$$= -\mathbb{E}\left[\log\left(\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=n}^{n} (y_i - A^T x_i)^2\right)\right)\right]$$

$$= -\mathbb{E}\left[\log\left(\frac{1}{(\sqrt{2\pi\sigma^2})^n}\right)\right] - \mathbb{E}\left[\frac{-1}{2\sigma^2} \sum_{i=1}^{n} (y_i - A^T x_i)^2\right]$$

$$= -\mathbb{E}\left[\log \frac{1}{(\sqrt{2\pi\sigma^2})^n}\right] + \frac{1}{2\sigma^2} \mathbb{E}\left[\sum_{i=1}^{n} (y_i - A^T x_i)^2\right]$$

$$= -\log \frac{1}{(\sqrt{2\pi\sigma^2})^n} + \frac{n}{2}$$

$$= \frac{n}{2} \log(2\pi\sigma^2) + \frac{n}{2}$$

Fisher information

- consider the case of an experiment making nanoparticles of two morphologies: sphere and cylindrical

- our model for a unit volume we produce number of particles in the following way.

$$n_{sph} = \alpha + \beta \quad \text{and} \quad \sigma_s \text{ variance}$$

$$n_{cyl} = \alpha \qquad\qquad \sigma_c \text{ variance}$$

- our goal is to predict how well an experimental design will be able to constrain the model parameters before doing the experiment

- Such that we can forecast the results of different volumes to measure and compare precision versus cost

- let us try to predict the output of our experiments based on some intuition.

  given data, we can infer n_{cyl}, n_{sph} by using the model above.

  if there are too many n_{cyl} (in the volume that we measured) then it would boost our estimate of α but depress the estimate β

- That means there is a covariance between our estimates of the two parameters

- variance in σ_s is the sum of variance in α and β while σ_c is equal to variance in α.

- The Fisher information matrix defined below makes this idea precise and quantitative

  we define our mode to be of N parameters $P_1, P_2, \ldots, P_N$ such that our observables are given by

$$f_b = f_b(P_1, P_2, \ldots, P_N)$$

  Fisher information matrix $F_{ij} = \sum_b \frac{1}{\sigma_b^2} \frac{\partial f_b}{\partial P_i} \frac{\partial f_b}{\partial P_j}$

- For the above simple model, we get

$$F = \begin{bmatrix} \frac{1}{\sigma_s^2} + \frac{1}{\sigma_c^2} & \frac{1}{\sigma_s^2} \\ \frac{1}{\sigma_s^2} & \frac{1}{\sigma_s^2} \end{bmatrix}$$

(columns labelled α, β; rows labelled α, β)

$P_1 = \alpha, \quad P_2 = \beta$

$n_s = f_1, \quad n_{cyl} = f_2$

$$F_{1,1} = \frac{1}{\sigma_s^2} \frac{\partial f_1}{\partial \alpha} \frac{\partial f_1}{\partial \alpha} + \frac{1}{\sigma_c^2} \frac{\partial f_2}{\partial \alpha} \frac{\partial f_2}{\partial \alpha}$$

$$= \frac{1}{\sigma_s^2}\left(1 + \frac{\partial \beta}{\partial \alpha}\right)^2 + \frac{1}{\sigma_c^2} = \frac{1}{\sigma_s^2} + \frac{1}{\sigma_c^2}$$

($\frac{\partial \beta}{\partial \alpha} \to 0$)

covariance of the model parameters is F^{-1}

$$= \begin{bmatrix} \sigma_c^2 & -\sigma_c^2 \\ -\sigma_c^2 & \sigma_s^2 + \sigma_c^2 \end{bmatrix}$$

- which verifies that α (# of cylinders in a vol.) is proportinoal to variance in cylinder production,

- one of the advantage of using Fisher information matricesis that we can encode prior knowledge by adding appropriate variance term to the matrix

  for example if we know that # of spherical particles in a given volum has the variance $\sigma_{s,p}^2$ then the fisher matrix should be modified as

$$\text{prior} = \begin{bmatrix} \sigma_{s,p}^2 & 0 \\ 0 & \sigma_{c,p}^2 \end{bmatrix}$$

$$I = \text{Prior}^{-1}$$

Thus $F_{prior} = F + \text{prior}^{-1}$

Optimal designs using Fisher information matrix:

- There are many ways to perform a optimal design that makes use of Fisher information matrix

- F measures the information that an observable random variable 'y' carries about an unknown parameter β such that

$$p(y) = f(y; \beta)$$

  i.e the variation in probability $p(y)$ is defined by β

- One such optimality criteria is D-optimality that maximizes the determinant of $F^T F$

  (For any matrix, the determinant computes the volume defined by cuboid it spans when applied to any vector of the corresponding)

  using this, we would say that D-optimality tries to maximize volume of information

  (Show example from dexpy)

- one algorithmic way of achieving optimal design is to use the coordinate-exchange algorithm.
  - at each iteration we randomly select two points from the candidate set, swap their values and see if the criteria improves. Repeat until we donot see any improvement

- The exchange step is primarily to improve exploration and avoid local optimum

- Another simple example would be measuring tensile strength by application of different levels of stress. Our regression model is more precise when the determinant of Fisher information matrix is large (low error of the model)