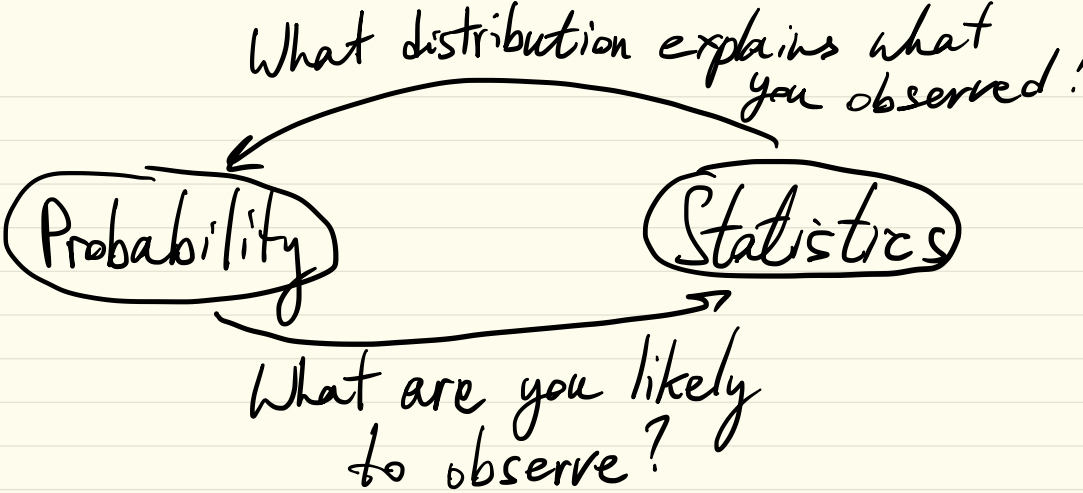
Lecture 14/15:

Chapter 3

Intro. to Statistical Inference

- Mean Square Error and the Bias-Variance decomposition
- Maximum Likelyhood Estimation



Basic Question:

Given a sample $x_1, \ldots, x_n \overset{iid}{\sim} P$, how do you infer properties of P?

Examples of Properties:

- mean, median, variance, CDF.

Two types of inference:

- parametric: the set of possible distributions is parametrized by some parameter in a nice space (maybe $\mathbb{R}^d$)

  Eg: Suppose $X_1, \ldots, X_n \sim N(\mu, \Sigma)$ (parameters)

- nonparametric: don't assume a "nice" parametrization.

Eg: Estimate the CDF of a R.V. X from iid realizations $X_1, \ldots, X_n$.

Suppose you want to estimate some parameter θ (e.g. mean, CDF)

An <u>estimator</u> $\hat{\theta}_n$ of θ is a function

$$\hat{\theta}_n = g(X_1, \ldots, X_n)$$

We say $\hat{\theta}_n$ is <u>unbiased</u>, if

$$\mathbb{E}\hat{\theta}_n = \theta$$

If $\hat{\theta}_n \in \mathbb{R}^d$, we can define the mean squared error:

$$MSE = \mathbb{E}\|\hat{\theta}_n - \theta\|_2^2$$

$$= \underbrace{\|\mathbb{E}\hat{\theta}_n - \theta\|_2^2}_{Bias^2(\hat{\theta}_n)} + \underbrace{\mathbb{E}\|\hat{\theta}_n - \mathbb{E}\hat{\theta}_n\|_2^2}_{Var(\hat{\theta}_n)}$$

Remarks:

- Tyically it pays to use biased estimators with lower variance. [More later]
- Ideally one would like confidence regions, e.g.

$$\mathbb{P}[\,\|\hat{\theta}_n - \theta\| \geq \delta_n\,] \leq 1-\alpha$$

- One often uses estimators to check hypothesis [E.g. is a coin biased?]

Maximum Likelihood Estimation

Suppose a parametric model

$$\{P_\theta\}_{\theta \in \Theta} = \{p(x;\theta) : \theta \in \Theta\}$$

densities or probability mass functions

The likelihood function is

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} p(x_i;\theta)$$

Log-likelihood function is

$$\ell_n(\theta) = \log \mathcal{L}_n(\theta)$$

The maximum likelihood estimator (MLE) is $\hat{\theta}_n \in \operatorname{argmax}_{\theta \in \Theta} \mathcal{L}_n(\theta)$

An interpretation based on KL divergence.

For two probability measure $P \ll Q$. define

$$D_{KL}(P \| Q) = \mathbb{E}_P \ln\left(\frac{dP}{dQ}\right)$$

Lemma: $D_{KL}(P \| Q) \geq 0$ with equality iff $P = Q$

[We'll prove a stronger result later]

Then given a parametric model we may wish to solve

$$\theta_{KL} := \underset{\theta \in \Theta}{\operatorname{argmin}} D_{KL}(p(x;\theta^*) \| p(x;\theta))$$

$$= \underset{\theta \in \Theta}{\operatorname{argmin}} \mathbb{E}_{p^*} \log\left(\frac{p(x;\theta^*)}{p(x;\theta)}\right)$$

So

$$\theta_{KL} := \underset{\theta \in \Theta}{\operatorname{argmax}} \mathbb{E}_{p^*} \log(p(x;\theta))$$

$$\approx \underset{\theta \in \Theta}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} \ell_n(\theta)$$

$$= \theta_{MLE}$$

Examples:

1) $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$

Then $p(x,p) = p^x (1-p)^{1-x}$ and

$$\mathcal{L}_n(p) = \prod_{i=1}^n f(x_i, p) = p^{S_n} (1-p)^{n-S_n}$$

where $S_n = \sum_{i=1}^n X_i$

So

$$\ell_n(p) = S_n \log p + (n - S_n) \log(1-p)$$

and

$$\hat{p}_n = \frac{S_n}{n} = \frac{\sum_{i=1}^n X_i}{n}$$

2) $X_1, \ldots, X_n \sim N(\mu, \Sigma)$

Then

$$\mathcal{L}_n(\mu, \Sigma) = (2\pi)^{-\frac{kn}{2}} \det(\Sigma)^{-\frac{n}{2}} \cdot \exp\left(-\frac{1}{2} \sum_{i=1}^n (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)\right)$$

So

$$\ell_n(\mu, \Sigma) = C - \frac{n}{2}\log(\det(\Sigma)) - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

concave in μ

$$(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) = \|\Sigma^{-1/2}(x_i-\mu)\|_2^2$$

Function is concave in μ.

So setting $\nabla_\mu = 0$ gives

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Then setting $\nabla_\Sigma = 0$ gives

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})(x_i-\hat{\mu})^T$$

[Check this!]

3) Regression:

Suppose

$$y_i = g(x_i) + \epsilon_i$$

where $x_i \overset{iid}{\sim} P$, $\epsilon_i \sim N(0, \sigma^2 I)$

Then $y_i \overset{ind}{\sim} N(g(x_i), \sigma^2)$

So

$$\mathcal{L}_n(g) = C \prod_{i=1}^{n} e^{-\frac{\|y_i - g(x_i)\|^2}{\sigma^2}}$$

So

$$\ell_n(g) = C - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \|y_i - g(x_i)\|^2$$

Linear Regression is when

$$\{g(x) = Ax + b : A \in \mathbb{R}^{d \times k}, b \in \mathbb{R}^k\}$$

4) Logistic Regression:

Observe $(x_i, y_i) \in \mathbb{R}^d \times \{-1, 1\}$

with

$$P(y_i = 1 | x) = S(g(x))$$

where

$$S(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x} = 1 - S(-x)$$

So

$$\mathcal{L}_n(g) = \prod_{i=1}^{n} S(g(x_i))^{\frac{1+y_i}{2}} \underbrace{(1 - S(g(x_i)))}_{S(-g(x_i))}^{\frac{1-y_i}{2}}$$

$$= \prod_{i=1}^{n} S(y_i g(x_i))$$

So

$$\ell_n(g) = \sum_{i=1}^{n} \ln\left(S(y_i g(x_i))\right)$$

$$= -\sum_{i=1}^{n} \ln\left(1 + \exp(-y_i g(x_i))\right)$$

When $\{g = a^T x + b : \begin{matrix} a \in \mathbb{R}^d \\ b \in \mathbb{R} \end{matrix}\}$

the MLE may be found with convex optimization

(More later)

5) Poisson Regression:

Observe $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{N}$

where

$$y_i | x \sim \text{Poiss}(g(x))$$

Then

$$p(y_i | x) = \frac{g(x)^{y_i} e^{-g(x)}}{y_i!}$$

So

$$\mathcal{L}_n(g) = \prod_{i=1}^{n} \frac{g(x_i)^{y_i} e^{-g(x_i)}}{y_i!}$$

$$\ell_n(g) = C + \sum_{i=1}^{n} y_i \ln(g(x_i)) - \sum_{i=1}^{n} g(x_i)$$

Typical Assumption is

$$\ln(g(x)) = a^T x + b$$

where $a \in \mathbb{R}^d$, $b \in \mathbb{R}$

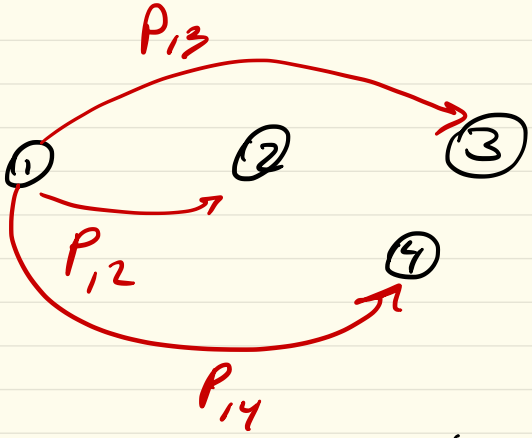
Then finding MLE becomes

$$\min_{a,b} \sum_{i=1}^{n} \exp(a^T x_i + b) - y_i (a^T x_i + b)$$

This is a convex problem.

6) Hidden Markov Model:

Suppose $X_1, \ldots, X_n$ is a markov chain on k nodes.



P_{ij} are transition probabilities.

[Exercise: What is the MLE for P from observations $X_1, \ldots, X_n$]

Suppose you observe only some y_i with density $p(y_i|x_i)$ and not x_i. Then

$$p(x_1, \ldots, x_n, y_1, \ldots, n)$$

depends on latent variables $x_1, \ldots, x_n$. We'll see later that finding MLE amounts to a nonconvex problem, but which can be approximately solved with Expectation Maximization.

Maximum likelihood can be understood in a broader context of stochastic optimization:

Goal: Solve $\min_{x \in \mathcal{X}} f(x) = \mathbb{E}_{z \sim P} \ell(x; z)$

Let x^* be a solution.

Two approaches:

1) Sample Average Approximation (SAA)

$$\hat{x}_n = \operatorname{argmin}_{x \in \mathcal{X}} f_n(x) := \frac{1}{n} \sum_{i=1}^{n} \ell(x; z_i)$$

2) Streaming Algorithms

{ Draw $z \sim P$
{ Update $x_{t+1} = \underset{\text{algorithm}}{A}(x_{1:t-1}, z_{1:t-1})$ }

The MLE falls within first class. We focus on ① now and will talk about ② later.

Two considerations:

① Assymptotic consistency $\hat{x}_n \xrightarrow{p} x^*$

② Assymptotic normality

$$\sqrt{n}(\hat{x}_n - x^*) \xrightarrow{d} N(0, \Sigma)$$

for some matrix Σ.

③ Finite sample bounds:

$$P(f(\hat{x}_n) - f(x^*) < \varepsilon) \geq \delta_n(\varepsilon)$$