Lecture 2

DJM

2 October 2018

Statistics vs. ML

- Lots of overlap, both try to "extract information from data"

Venn diagram

Probability

1. X_n converges *in probability* to X, $X_n \xrightarrow{P} X$, if for every $\epsilon > 0$, $\mathbb{P}(|X_n - X| < \epsilon) \to 1$.
2. X_n converges *in distribution* to X, $X_n \rightsquigarrow X$, if $F_n(t) \to F(t)$ at all continuity points t.
3. (Weak law) If $X_1, X_2, \ldots$ are iid random variables with common mean m,then $\overline{X}_n \xrightarrow{P} m$.
4. (CLT) If $X_1, X_2, \ldots$ are iid random variables with common mean m and variance $s^2 < \infty$, then $\sqrt{n}(\overline{X}_n - m)/s \rightsquigarrow \mathrm{N}(0, 1)$.

Big-Oh and Little-Oh

Deterministic:

1. $a_n = o(1)$ means $a_n \to 0$ as $n \to \infty$
2. $a_n = o(b_n)$ means $\frac{a_n}{b_n} = o(1)$.
   Examples:
   - If $a_n = \frac{1}{n}$, then $a_n = o(1)$
   - If $b_n = \frac{1}{\sqrt{n}}$, then $a_n = o(b_n)$
3. $a_n = O(1)$ means a_n is eventually bounded for all n large enough, $|a_n| < c$ for some $c > 0$. Note that $a_n = o(1)$ implies $a_n = O(1)$
4. $a_n = O(b_n)$ means $\frac{a_n}{b_n} = O(1)$. Likewise, $a_n = o(b_n)$ implies $a_n = O(b_n)$. Examples:
   - If $a_n = \frac{n}{2}$, then $a_n = O(n)$

Stochastic analogues:

1. $Y_n = o_p(1)$ if for all $\epsilon > 0$, then $P(|Y_n| > \epsilon) \to 0$
2. We say $Y_n = o_p(a_n)$ if $\frac{Y_n}{a_n} = o_p(1)$
3. $Y_n = O_p(1)$ if for all $\epsilon > 0$, there exists a $c > 0$ such that $P(|Y_n| > c) < \epsilon$
4. We say $Y_n = O_p(a_n)$ if $\frac{Y_n}{a_n} = O_p(1)$
   Examples:
   - $\overline{X}_n - \mu = o_p(1)$ and $S_n - \sigma^2 = o_p(1)$. By the the Law of Large Numbers.
   - $\sqrt{n}(\overline{X}_n - \mu) = O_p(1)$ and $\overline{X}_n - \mu = O_p(\frac{1}{\sqrt{n}})$. By the Central Limit Theorem.

Statistical models

A statistical model $\mathcal{P}$ is a collection of probability distributions or densities. A parametric model has the form

$$\mathcal{P} = \{p(x; \theta) : \theta \in \Theta\}$$

where $\Theta \subset \mathbb{R}^d$ in the parametric case.

Examples of nonparametric statistical models:

- $\mathcal{P} = \{$ all continuous CDF's $\}$
- $\mathcal{P} = \{f : \int (f''(x))^2 dx < \infty\}$

Evaluating estimators

An *estimator* is a function of data that does not depend on θ.

Suppose $X \sim N(\mu, 1)$.
-μ is not an estimator.
-Things that are estimators: X, any functions of X, 3, $\sqrt{X}$, etc.

1. Bias and Variance
2. Mean Squared Error
3. Minimaxity and Decision Theory
4. Large Sample Evaluations

MSE

Mean Squared Error (MSE). Suppose $\theta, \widehat{\theta}$, define

$$\mathbb{E}\left[\left(\theta - \widehat{\theta}\right)^2\right] = \int \cdots \int \left[\left(\widehat{\theta}(x_1, \ldots, x_n) - \theta\right) f(x_1; \theta)^2 \cdots f(x_n; \theta)\right] dx_1 \cdots dx_n.$$

Bias and Variance The bias is

$$B = \mathbb{E}\left[\widehat{\theta}\right] - \theta,$$

and variance is

$$V = \mathbb{V} \| \left[\widehat{\theta}\right].$$

Bias-Variance Decomposition

$$MSE = B^2 + V$$

$$\begin{aligned}
MSE &= \mathbb{E}\left[(\widehat{\theta} - \theta)^2\right] \\
&= \mathbb{E}\left[\left(\widehat{\theta} - \mathbb{E}\left[\widehat{\theta}\right] + \mathbb{E}\left[\widehat{\theta}\right] - \theta\right)^2\right] \\
&= \mathbb{E}\left[\widehat{\theta} - \mathbb{E}\left[\widehat{\theta}\right]\right] + \left(\mathbb{E}\left[\widehat{\theta}\right] - \theta\right)^2 + \underbrace{2\mathbb{E}\left[\widehat{\theta} - \mathbb{E}\left[\widehat{\theta}\right]\right]}_{=0} \left(\mathbb{E}\left[\widehat{\theta}\right] - \theta\right) \\
&= V + B^2
\end{aligned}$$

An estimator is unbiased if $B = 0$. Then $MSE = Variance$.

Let $x_1, \ldots, x_n \overset{iid}{\sim} N(\mu, \sigma^2)$.

$$\begin{aligned}
\mathbb{E}[\overline{x}] &= \mu, & \mathbb{E}\left[s^2\right] &= \sigma^2 \\
\mathbb{E}\left[(\overline{x} - \mu)^2\right] &= \frac{\sigma^2}{n} = O\left(\frac{1}{n}\right) & \mathbb{E}\left[(s^2 - \sigma^2)^2\right] &= \frac{2\sigma^4}{n-1} = O\left(\frac{1}{n}\right).
\end{aligned}$$

Minimaxity

Let $\mathcal{P}$ be a set of distributions. Let θ be a parameter and let $L(\theta, \theta')$ be a loss function.

The **minimax risk** is

$$R_n(\mathcal{P}) = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[L(\theta, \widehat{\theta})]$$

If $\sup_{P \in \mathcal{P}} \mathbb{E}_P[L(\theta, \widehat{\theta})] = R_n$ then $\widehat{\theta}$ is a minimax estimator.