Lecture 6 - scribbles

Today: finish decision theory

gen/disc. learning

Start decision theory cold.

Given previous framework, how can we compare stat procedures? E.g., \( \mathcal{S}_1 \) vs. \( \mathcal{S}_2 \)

First property

(frequentist) risk

\[
R(P, S) = \mathbb{E}_{\mathcal{D}} \left[ L(P, S(\mathcal{D})) \right]
\]

"Risk profiles"

\[
\begin{array}{c}
R(\theta, S) \quad S_1 \\
\downarrow \quad \downarrow \\
S_0 \quad \theta
\end{array}
\]

**Transform to scalars**

- "Minimax" analysis: \( \max_{\mathcal{P} \in \mathcal{B}} \max_{\theta} \mathbb{E}_{\theta} R(\theta, S) \)
- Frequentist
- Weighted average: \( \int_R R(\theta, S) \pi(\theta) d\theta \)

"Worst case"
* Weighted average: \( \int \mathbb{R}(\theta, S) \pi(\theta) \ d\theta \)

* Alternatively, in ML theory, is PAC theory

\[ \text{look at small bounds for dist. of } L(P, S(D)) \quad [D \text{ is random}] \]

\[ \text{PAC band} \]

\[ \mathbb{P} L(P, S(D)) > \text{stuff} \quad \text{if small number} \]

"with high prob" statement

Bayesian decision theory:

\( R_b(a | D) = \int_{\theta} L(\theta, a) p(\theta | D) \ d\theta \)

Bayesian optimal action:

\( S_{\text{bayes}}(D) = \arg\max_{a \in A} R_b(a | D) \)
Example: if $f = \Theta$ ("estimation")

$L(\Theta, \alpha) = (10 - \alpha)^2$

Then (exercise) $\text{Bayes}(D) = E[\Theta|D]$ (posterior mean)

\begin{itemize}
  \item \textbf{Bayesian posterior risk}
  \item $f_D = P(\Theta|D)$ - marginal dist
  \item $p(D) = \int \text{Bayesian posterior } f_D \text{ minimizes this when using a prior } p(\Theta) = \pi(\Theta)$
\end{itemize}

Examples of estimators: \(s: D \rightarrow \Theta\)

1) MLE
2) MAP
3) method of moments

\text{idea: find an injective mapping from } \Theta \text{ to 'moments'} \frac{EX}{EX^2}$
and subjective on "possible moments"
and then \( \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \)
\( \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \)

\[ f(\mu, \sigma^2) \triangleq \left( \frac{\hat{\mu}}{\hat{\sigma}^2} \right) \]
\[ \left( \frac{\hat{\mu}}{\hat{\sigma}^2} \right) \triangleq f^{-1} \left( \frac{\hat{\mu}}{\hat{\sigma}^2} \right) \]

(here, this estimator is same as MLE)
[property of exponential family]

\( \circ \) this is useful for latent variable models (e.g. mixture of Gaussians)
(spectral methods) e.g.

4) in the context of prediction \( \mathcal{A} = \mathbb{E} (f: \mathcal{X} \rightarrow \mathcal{S}) \)
\( \mathcal{X} \) input space
\( \mathcal{S} \) output

example of \( f: \mathcal{X} \rightarrow \mathcal{A} \)
is using empirical "risk" minimization (ERM)

Vapnik risk:

\[ L(p, f) = \mathbb{E}_{(y, x) \sim p} [l(y, f(x))] \]

replace with

\[ \mathbb{E}_{(y, f(x)) \sim \mathcal{F}} = \frac{1}{n} \sum_{i=1}^{n} l(y_i, f(x_i)) \]

\[ \frac{1}{n} = \text{average } \mathbb{E}_{(y, f(x)) \sim \mathcal{F}} \]

hypothesis class

\[ \text{bias-variance decomposition} \]

i.i.d. setting \( D = \{ x \}_{i=1}^{n} \)

\( D_n = \{ x_i \}_{i=1}^{n} \)

\( x_i \) i.i.d. \( p \)

notation: \( \hat{\epsilon}_n = \hat{\epsilon}_n (D_n) \)

\( \text{to highlight dependence on } n \)

study \( R(e, \hat{\epsilon}_n) \) as a function of \( n \)

in particular, would like \( R(e, \hat{\epsilon}_n) \rightarrow R(e, \tilde{\epsilon}_n) \)

as \( n \rightarrow \infty \)
in particular, would like $K(\theta, \mathcal{D}_n) \to K(\theta, \mathcal{D})$

"consistency" argmin $K(\theta, \mathcal{D})$

* for estimation, typical loss: squared loss $L(\theta, \mathcal{S}_n(\theta)) = ||\theta - \mathcal{S}_n(\theta)||^2$

standard statistical consistency

\[ \hat{\theta}_n \overset{P}{\to} \theta \]

"in probability"

\[ \text{i.e. } \forall \varepsilon > 0, \quad \Pr[|\hat{\theta}_n - \theta| > \varepsilon] \to 0 \]

\[ \text{randomness is from } \mathcal{D}_n \]

\[ \hat{\theta}_n = \mathcal{S}_n(\mathcal{D}_n) \]

* last time, $R(\theta, \mathcal{S}_n) = E_{\mathcal{D}_n}[||\theta - \hat{\theta}_n||^2]$

\[ = ||\theta - E\hat{\theta}_n||^2 + \text{bias}^2 + E[||\hat{\theta}_n - E\hat{\theta}_n||^2] \]

James-Stein estimator:

for estimating the mean of $N(\mu, \sigma^2 I)$

\[ S_{JS} \text{ is biased, but lower variance than MLE} \]

\[ \text{So, actual simulation is x - x} \]
\( \alpha \) is \underline{poorer}, but lower variance than MLE

\( S \)s actually \underline{strictly dominates} \underline{MLE} \( \theta \rightarrow 3 \)

i.e. \( R(\theta, S) \leq R(\theta, \text{MLE}) \) for \( \theta \)

and \( \exists \theta \) s.t. \( R(\theta, S) < R(\theta, \text{MLE}) \)

\( \Rightarrow \) MLE is sometimes \underline{inadmissible}.

---

In assignment, "consistency" mean \( R(\theta, S_n) \overset{p}{\rightarrow} 0 \)

(\( \overset{p}{\rightarrow} \) \( \Rightarrow \) \( \overset{\text{convergence in } P}{} \))

"convergence in \( P \)"

by \underline{bias-variance decomposition}:

\[
\text{bias}(S_n) \overset{n \rightarrow \infty}{\rightarrow} 0 \\
\text{variance}(S_n) \overset{n \rightarrow \infty}{\rightarrow} 0
\]

\( \Rightarrow R(\theta, S_n) \overset{n \rightarrow \infty}{\rightarrow} 0 \) \( \Rightarrow \) consistency

\( \overset{\text{in probability}}{} \)

properties (asymptotic) of MLE:

\( \text{under regularity conditions on } \Theta \subseteq \mathbb{P}(x \in \Theta) \)

a) \( \hat{\theta}_n \overset{P}{\rightarrow} \theta \) "consistent" in distribution

b) CLT: \( \sqrt{n}(\hat{\theta}_n - \theta) \overset{d}{\rightarrow} N(0, \text{var}(\theta)) \)

\( \text{Central Limit Theorem } \)

\( \text{under id. data conditions } \)

\( \text{information matrix } \)
Theorem (Cramer-Rao Lower Bound)

C) asymptotically optimal

ie. it has minimal asymptotic variance among all "reasonable" estimators

Let \( \hat{\theta} \) be consistent

D) Invariance: MLE is preserved under reparameterization

Suppose there is a bijection \( f: \Theta \rightarrow \Theta' \)

\[ \hat{\theta} \rightarrow \hat{\theta}' \]

then \( f(\hat{\theta}) = \hat{\theta}' \)

If not a bijection, can generally MLE with "profile likelihood"

Suppose \( g: \Theta \rightarrow \Delta \) profile likelihood \( L(\theta) \triangleq \max_{\theta \in \Delta} \rho(\text{data}; \theta) \)

\( \theta \in \Theta' \)

Define \( \hat{\theta}_{MLE} = \arg\max_{\theta \in \Delta} L(\theta) \)

then we have \( \hat{\theta}_{MLE} = g(\hat{\theta}_{MLE}) \)

"Plug in" estimator

Example:

\[ (\hat{\sigma}^2) = (\hat{\sigma}')^2 \]

\[ \sin \sigma^2 = \sin \hat{\sigma}^2 \]