Today:

- MLE chart

Statistical decision theory

Optimization comments about MLE:

\[ \min_{\theta \in \Theta} f(\theta) \]

\( \text{if } f(\theta) \text{ is differentiable, } \nabla f(\theta^*) = 0 \) (stationary points)

For \( \theta^* \) to be a local min when \( \Theta^* \) is in interior of \( \Theta \):

- Also check that Hessian \( (f)(\theta^*) \) is positive definite for a local min.

\[ H > 0 \iff \nabla^2 f(\theta^*) > 0 \quad \forall u \in \mathbb{R}^d \]

\[ (f)(\theta^*) > 0 \]

- Only local results in general.

But if Hessian \( (f)(\theta) \) > 0 \( \forall \theta \in \Theta \), \( \theta^* \) is said to be

then \( \nabla f(\theta^*) = 0 \Rightarrow \theta^* \) is a global min.

- Otherwise, for smooth func., looking at all gradient pts. and boundary pts.

- Give you enough information to find global min.

Be careful with boundary cases:

\( f(\theta) \neq 0 \)

\( \Theta \) boundary case, e.g.

Then example:

\( \theta^* \rightarrow \infty \)

Example where MLE does not exist:

- Does not always exist [\( \Theta \in \mathbb{R} \) but \( \Theta \) is open] or when \( \theta^* = +\infty \)

\[ \Theta = [0, 1] \]
Example II: multinomial distribution

Suppose \( X_i \) is a discrete r.v. on \( k \) choices "multinomial" (we could choose \( \mathcal{X}_i = \{1, 2, \ldots, k\} \))

but instead, convenient to encode the \( k \) possibilities using unit boxes in \( \mathbb{R}^k \).

i.e. \( \mathcal{X}_i = \{e_1, \ldots, e_k\} \) where \( e_j \in \mathbb{R}^k \) "one hot encoding"

\[ e_j = \left( \frac{1}{k}, \ldots, \frac{1}{k}, 0 \right) \] in coordinates

Parameter for discrete r.v. \( \pi \in \Delta_k \) (where \( \mathcal{X}_i = \{e_1, \ldots, e_k\} \))

\[ \Delta_k = \{ \pi \in \mathbb{R}^k : \pi_j \geq 0, \sum_j \pi_j = 1 \} \]

Probability simplex on \( k \) choices

We will write \( X_i \sim \text{Mult}(\pi) = \text{Mult}(n, \pi) \)

\( \mathbb{E} \) consider \( X_i \sim \text{Mult}(\pi) \)

then \( \mathbf{X} = \frac{X}{n} \sim \text{Mult}(n, \pi) \)

"multinomial distribution"

\( \mathbf{X} \in \mathbb{N}^k \)

\[ \sum_{j=1}^k n_j = n \]

Proof for \( \mathbf{X} : \)

\[ p(\mathbf{x}|\pi) = \binom{n}{n_1, \ldots, n_k} \frac{n!}{n_1! \cdots n_k!} \frac{\pi_1^{x_1} \cdots \pi_k^{x_k}}{\pi_1 \cdots \pi_k} \]

\( \leq \) multinomial coeff

\[ \mathbf{x} = (n_1, \ldots, n_k) \]

\( \sum_{i=1}^k n_i \leq n \)

\[ \frac{n!}{(n_1 \cdots n_k)!} \]

\[ \binom{n}{n_1, \ldots, n_k} = \frac{n!}{n_1! \cdots n_k!} \]

15/29
multinomial MLE

\[
\log\text{-likelihood} \quad l(\pi) = \log p(x | \pi) = \log \left( \frac{n}{n_1 \cdots n_k} \right) + \sum_{j=1}^{k} n_j \log \pi_j
\]

\[\pi = (n_1, \ldots, n_k)\]

MLE:

\[
\hat{\pi}_{\text{MLE}}(x) = \arg\max_{\pi \in \Delta_k} l(\pi)
\]

\[\text{subject to } \sum_{j=1}^{k} \pi_j = 1\]

too many parameters

a) reparameterize problem so that \( \beta \) is full dimensional

\[
\tau_j = 1 - \frac{\pi_j}{\sum_{i=1}^{k} \pi_i}
\]

\[\pi_1, \ldots, \pi_k \in [0,1] \text{ with constraint } \sum_{j=1}^{k} \pi_j = 1\]

how magic that \( \log \pi_j \) acts as a barrier far away from \( \pi_j = 0 \)

can do unconstrained optimization on \( \tau_1, \ldots, \tau_k \)

of \( l(\tau_1, \ldots, \tau_k) \)

having set \( n \) is in the interior of constraint set

and it usually well for log-type problems

b) use Lagrange multiplier approach to handle equality constraint on \( \Delta_k \)

\[
\max f(\pi)
\]

\[\text{subject to } g(\pi) = 0
\]

\[\sum_{j=1}^{k} \pi_j = 1
\]

\[\pi_j \geq 0
\]

Lagrange multiplier method is look at stationary pts. of \( J(\pi, \lambda) \)

(gradient)

\[\nabla J(\pi, \lambda) = 0\]

necessary cond. for local opt.

\[\pi, \lambda \text{ are stationary}
\]

(check "bottled Heron" to get local max or min)
$l(x) = \sum \log \pi_j$

(strictly concave

(at ln $\pi_j$)

$\frac{\partial l}{\partial \pi_j} = \lambda = 0 \Rightarrow \pi_j^* = \frac{n_j}{n}$

$\frac{\partial l}{\partial \pi_j} = \lambda \Rightarrow \pi_j^* = \frac{n_j - 1}{n - 1}$

$\Rightarrow \pi_j^* = \frac{n_j}{n}$

notice $\pi_j^* = \frac{n_j}{n} \in \text{dom}$

$\nabla \pi L(\pi, \lambda) = 0 \Rightarrow \nabla \pi f(\pi) + \lambda \nabla g(\pi) = 0$

$\Rightarrow \nabla f(\pi) = -\lambda \nabla g(\pi)$

Statistical decision theory

A) Bias-variance decomposition for squared loss

$E_{\text{MLE}}(x) = \arg \max_{\theta \in \Theta} p(x | \theta)$

$E_{\text{MAP}}(x) = \arg \max_{\theta \in \Theta} p(x | \theta) p(\theta)$

$\Rightarrow \ln p(x) = \ln p(x | \theta) + \ln p(\theta)$

$\times$ how do we evaluate these estimates?

$\hat{S} : \Omega \rightarrow \Theta$

$\hat{S} = S(X)$

most standard tool: $\text{expected risk of an estimator}$

$R(\hat{S}, \theta) \equiv E_{\theta} [L(\hat{S}, S(x))]$
Squared loss: \( L(\theta, \tilde{\theta}) = \| \theta - \tilde{\theta} \|_2^2 \quad \tilde{\theta} = \bar{y}(x) \)

\[
E_X \left[ \| \theta - \tilde{\theta} \|_2^2 \right] = E \left[ \| \theta - \tilde{\theta} \|_2^2 \right] + E \left[ \| \theta - \bar{y} \|_2^2 \right] + 2 \left< \theta - \bar{y}, \tilde{\theta} - \bar{y} \right>
\]

\[
= \| \theta - \bar{y} \|_2^2 \quad \text{variance}\]

\[
\text{bias}^2 = \| \theta - \bar{y} \|_2^2
\]

\[
\text{(freq) risk for squared loss} = \text{bias}^2 + \text{variance}
\]

\*(frequency) risk for squared loss = bias\(^2\) + variance

* Consistency: informally "do right thing as \( n \to \infty \)" where \( n \) is training set size

\[
X \to (x_i) \overset{n}{\longrightarrow} \:
\hat{\theta}_n \quad (\text{data of size} \ n)
\]

\[
\text{Consistency: } \text{bias}(\hat{\theta}_n) \to 0 \:
\Rightarrow \quad \text{variance}(\hat{\theta}_n) \to 0 \Rightarrow \quad \hat{\theta}_n \text{ is consistent}
\]

\[
\hat{\theta}_n \to \theta \quad \text{(bias)}
\]

\[
\text{bias}(\hat{\theta}_n) \to 0
\]

\[
\text{variance}(\hat{\theta}_n) \to 0
\]

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