# Lecture 14

LVII posted

CP recap

Missing data

403 Course
# Overview

- EDF $\hat{F}$ and consistency of means
- MC integration ← Application
- How to sample
  - $F^{-1}$
  - importance s.
  - rejection s.
- * (CV)
- * Bootstrap ← $Var(\hat{\theta}) \approx ?$
- * Conformal prediction ✓ ←
- * Imputation (fill missing data) ←
- MCMC (Markov Chain MC)

Resampling

# Conformal prediction

new x

AFTER TRAINING

- **Conformal prediction:** CI for a single prediction $\hat{y} = f(x)$

Given data set $\mathcal{D} = \{(x_i, y_i), i = 1 : N\}$
Training algorithm $\mathcal{A}$, so that $\mathcal{A}(\mathcal{D}) = f$ the predictor
Want CI for $\hat{y} = f(x)$ where $x$ is a new data point
so that the CI is NOT dependent on $\mathcal{A}$ being statistically correct (e.g. $\mathcal{A}$ overfits, ...)

GOOD !!

- `jackknife+` is a simple algorithm for CP
- More advanced algorithms exist. This is an active area of research in statistics.

!!!! Do NOT use CP to "crossvalidate" your algorithm!

jackknife+

$CI_\alpha = [a, b]$

MUST contain ALL [Training, CV, Model selection, ...

new x:

$x$

$\mathcal{D} \to \boxed{A} \to f \to f(x) = \hat{y} \leftarrow$ prediction

using $f$

$\mathcal{D}_{-i} \to \boxed{A} \to f_{-i} \to f_{-i}(x)$, $R_i = |y^i - f_{-i}(x^i)|$

$i = 1:n$

$x$

$f_{-i}(x) - R_i = a_i$

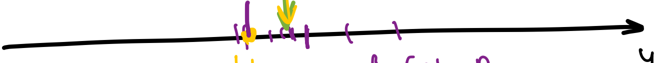
$\alpha = 20\%$

$\frac{\alpha}{2} = 10\%$

$a = \lfloor n\frac{\alpha}{2} \rfloor$ quantile $\to \{a_i, i = 1:n\}$

a

$f(x)$ prediction



$y$

skip $\lfloor n\frac{\alpha}{2} \rfloor$

$a_1 = f_{-1}(x) - R_1$

# Lecture Notes VII – Missing Data and Imputation

" fill in

Marina Meila
mmp@stat.washington.edu

Department of Statistics
University of Washington

April 2025

The missing data problem

# Missing data

data $y_{1,2,\ldots N}$ some missing values

$\boxed{X_{1:n}}$ $y_{1:n}$ observed
$\boxed{X_{n+1:N}}$ $y_{n+1:N}$ unobserved

} want $\hat{\theta}$ param of interest

other information, known

$y_{n+1:n+m}$ = votes to be removed
impute
$\boxed{X_{n+1:n+m}}$ = addresses
impute
$Z_{n+1:n+m}$ = gender

## Missing data: the problem

- Data sampled i.i.d. $\sim F$, but some (parts of each sample) are missing
- Here we denote data point $= (X^i, Y^i)$ where $X^i$ is always observed, $Y^i$ may be missing or not.

complete

$R_i = \begin{cases} 1 & Y_i \text{ observed} \\ 0 & \text{otherwise} \end{cases}$

- A new random variable $R$, $R^i = 0$ when $Y^i$ not observed, $R^i = 1$ when observed
- **What is the distribution of $R$?** dependence of $R$ on $X, Y$

MCAR Missing Completely At Random $R \perp X, Y$

MAR Missing At Random $R \perp Y \mid X$

A, B events, r.v.'s

$\boxed{A \perp B}$

A independent of B

$P(A,B) = P(A)P(B)$

$P(A|B) = P(A)$

$P(B|A) = P(B)$

? know A

MCAR $\quad Y_{n+1:m} = ?$

MAR $\quad R \perp Y \mid X$

and R depends on X, but NOT on Y

## What is the distribution of $R$? how does R depend on X and Y?

- Let data $\mathcal{D} = \{(x^1, y^1), \ldots (x^n, y^n)\} \cup \{x_{n+1}, \ldots x_{n+m}\}$
  - complete data $\{(x^1, y^1), \ldots (x^n, y^n)\}$ with $R^{1:n} = 1$
  - data with missing values $\{x_{n+1}, \ldots x_{n+m}\}$ with $R^{n+1:n+m} = 0$

MCAR Missing Completely At Random $R \perp X, Y$

- $\{(x^1, y^1), \ldots (x^n, y^n)\}$ has no information about $\{y^{n+1}, \ldots y^{n+m}\}$

MAR Missing At Random $R \perp Y \mid X$

- $\{(x^1, y^1, R^1), \ldots (x^n, y^n, R^n)\}$ together with $\{(x^{n+1}, R^{n+1}), \ldots (x^{n+m}, R^{n+m})\}$ has information about $\{y^{n+1}, \ldots y^{n+m}\}$
- Probability distribution of missing values $Y^{n+1:n+m}$ is

estimate this!

$$Y \sim p(Y \mid X, R = 0) \quad \text{and} \quad \boxed{p(Y \mid X, R = 0) = p(Y \mid X, R = 1)} \tag{1}$$

use it to predict Y

$R \perp Y \mid X \iff P(Y \mid X, R) = P(Y \mid X)$

MAR $R \perp Y \mid X$

and R depends on X, but NOT on Y