MA 6973 Computational Statistics Sep 8, 2021

Logistics:
- Attendance
- 4 homework, mix of written & computational (returned via Gradescope)
  - ↳ (Python, Julia, Matlab).
    - ↳ Stan
- In-class midterm
- Final presentation & project, groups of 2 or 3. (+ report, in LaTeX).
  - ↳ overleaf
- Course website

Other
- Campuswire?
-

Overview of course:

1. "Classical numerical analysis"
   - ↳ root finding of function $f(x) = 0$
   - ↳ optimization
     - ↳ think finding MLE's, or fitting a NN
   - ↳ numerical linear algebra
     - ↳ solving $A\vec{x} = \vec{b}$
     - ↳ finding eigenvalues.
     - ↳ various factorizations, identities.
   - ↳ function approximation (generally, not interpolation)
     - least squares
     - regression
     - density estimation

$$\begin{cases} f_1(x_1, \dots, x_n) = 0 \\ \vdots \end{cases}$$

↳ integration

Compute $\int f(x)\,dx$ numerically

Ex: one might want to compute marginal distributions $p(x) = \int p(x,y,z)\,dy\,dz$

2. statistics, but methods too complicated for pen + paper.

- "simulation": - random variable generation (including MCMC)
  - estimate confidence intervals
  - regression (linear / nonlinear → Gaussian Process reg., etc.)
  - density estimation
  - conditional, hierarchical modelling (large scale Bayes).

Example CERN statistical model

Probability Topics Review

see my old webpage
cims.nyu.edu/~oneil/prob20

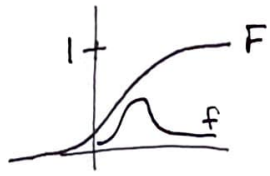
Random Variables: Discrete vs. Continuous

(Refs: Durrett or Ross)

less focus on these in this course

each R.V. has a pdf and distribution function

$$F(x) = \int_{-\infty}^{x} f(u)\,du$$



X is r.v., we say $X \sim \text{Normal}(\mu, \sigma^2)$

$$\Rightarrow \quad f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

for example.

(e.g. exponential, uniform, Gamma, etc. Cauchy)

$\Rightarrow$ Expected value, variance

$$\mathbb{E}(X) = \int x\, f(x)\, dx$$

$$\text{Var}(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right)$$

$\Rightarrow$ Conditional probabilities, independence,

Bayes. (for events)

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$ Conditional probability

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$$ Independence.

$$\Rightarrow \mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\,\mathbb{P}(B)$$

$$= \mathbb{P}(B \mid A)\,\mathbb{P}(A)$$

$$\Rightarrow \mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\,\mathbb{P}(A)}{\mathbb{P}(B)}$$

$$= \frac{\mathbb{P}(B \mid A)\,\mathbb{P}(A)}{\mathbb{P}(B \mid A)\mathbb{P}(A) + \mathbb{P}(B \mid A^c)\,\mathbb{P}(A^c)}$$ Bayes.

Functions of R.V.s

If X has pdf / dist f, F, and

$Y = \varphi(X)$, what is pdf/dist for Y?

$$\Rightarrow G(y) = \mathbb{P}(Y \le y) = \mathbb{P}(\varphi(X) \le y) = \mathbb{P}\left(X \le \varphi^{-1}(y)\right)$$

$$= \int_{x \le \varphi^{-1}(y)} f(x)\,dx \quad \ldots$$

Multiple R.V.s - Joint distributions

X has pdf f,

Y has pdf g,

X, Y has joint pdf $h(x,y)$:

$$\Rightarrow \mathbb{P}(X \in A, Y \in B) = \int_B \int_A h(x,y)\, dx\, dy$$

$$\mathbb{P}(X, Y \in \omega) = \iint_\omega h(x,y)\, dx\, dy$$

Conditional density: $p(x|y) = \dfrac{h(x,y)}{g(y)}$

$$\mathbb{P}(X \in A \mid Y = y) = \int_A p(x|y)\, dx$$

Marginal densities: $f(x) = \int h(x,y)\, dy$

$g(y) = \int h(x,y)\, dx$.

Sums:

$Z = X + Y$

what is the pdf of Z?

$$p(z) = \int_x f(z-x)\, f_Y(x)\, dx$$

Change of variables

$U = g(X,Y)$

$V = h(X,Y)$

what is pdf $f(u,v)$?

Covariance

$$\mathrm{Cov}(X,Y) = \mathbb{E}\left((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))\right)$$

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

Multivariate Normal Random Variables — special, important dist.

$$\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}, C \right) \qquad C_{ij} = \mathrm{Cov}(X_i, X_j)$$

- C is an $n \times n$ matrix, symmetric, positive semi-definite

$\downarrow$

$\lambda_i \geq 0$

pdf is given by:

$$f(x_1,\dots,x_n) = \frac{1}{(2\pi)^{n/2}\sqrt{\det C}}\, e^{-(\vec{x}-\vec{\mu})^T C^{-1}(\vec{x}-\vec{\mu})}$$

linear algebra!

$$\vec{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}, \quad \vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

If A is a $k \times n$ matrix, then

$\vec{Y} = A\vec{X}$ is also multivariate normal, and

$\mathbb{E}(\vec{Y}) = A\vec{\mu}$, and $\mathrm{Cov}(\vec{Y}) = ACA^T$

} normal r.v. invariance under linear transformation.

Limit Theorems , inequalities

(Notation of Convergence) come back to in stats review.

WLLN:

$\mu = \mathbb{E}(X_i) < \infty$

$\sigma^2 = \mathrm{Var}(X_i) < \infty$

Then for any $\epsilon > 0$

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_1^n X_n - \mu \right| \geq \epsilon \right) \to 0 \quad \text{as} \quad n \to \infty.$$

CLT

Same assumptions (other assumptions possible)

$$\frac{\frac{1}{n}\sum_1^n X_i - \mu}{\sigma/\sqrt{n}} \rightsquigarrow N(0,1)$$

$\rightsquigarrow$: convergence in distribution

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\frac{1}{n}\sum_1^n X_i - \mu}{\sigma/\sqrt{n}} \leq z \right) = \Phi(z).$$

SLLN

$\mu = \mathbb{E}(X_i) < \infty$

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{1}{n} \sum_1^n X_i = \mu \right) = 1.$$

Statistics

Refs: Wasserman, All of Statistics

Rice, Mathematical Statistics & Data Analysis

Casella & Berger, Statistical Inference

Old course website: cims.nyu.edu/~oneil/stat21

- Inference
- Hyp. testing
- Bayesian methods (inferen
- Regression, "curve fitting"
- Density estimation
- Sampling methods (MCMC).

Parametric Methods

Given data $x_1, \ldots, x_n$ (IID) assumed to be from distribution with pdf $f(x;\theta)$, infer θ. This is inference...

Maximum Likelihood

$$\mathcal{L}(\theta) = \text{joint density evaluated at the data}$$

$$= \prod_i f(x_i;\theta) \leftarrow \text{a function of } \theta.$$

Log-likelihood:

$$\ell(\theta) = \log \mathcal{L}(\theta)$$

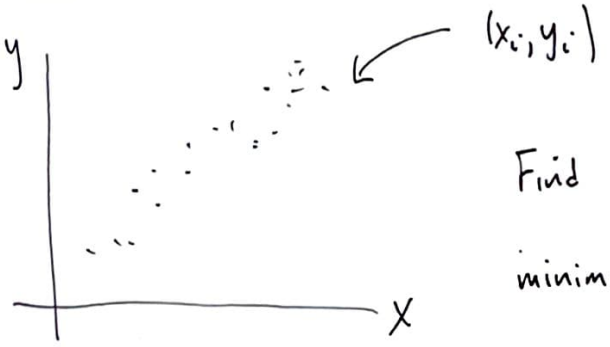
$$= \sum \log f(x_i;\theta)$$

$\hat{\theta}$ = MLE is solution to $\ell'(\theta) = 0$.

How to study the behavior of $\hat{\theta}$? Assume x_i are random variables...

How to solve $\ell'(\theta) = 0$ numerically?

Linear Regression



Find line $\hat{y} = \hat{a}x + \hat{b}$ that minimizes "something".

Model: $y = ax + b + \epsilon$

ϵ: IID with $\text{var}\,\epsilon = \sigma^2$, $E = 0$

One option: least squares

Minimize $\sum_i (y_i - \hat{y}_i)^2 = RSS$

Interpretation:
1. linear algebra
2. statistics (MLE, etc.)
3. calculus — take derivatives, solve optimization problem

solve ~~$A \begin{pmatrix} a \\ b \end{pmatrix} = A^T y$~~ $A^T A \begin{pmatrix} a \\ b \end{pmatrix} = A^T \hat{y}$

normal equations

when $A = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}$

What is the most accurate way to solve this linear algebra problem?

Other types of "data fitting":
- use basis functions $\varphi_1, \ldots, \varphi_n$
- Gaussian process regression
- "smoothing" ← usually non-parametric

Bayesian Methods

Goal: provide density for "belief" on value of a parameter θ
(can get philosophical...)

Idea: Put prior belief on θ, update to posterior belief by incorporating the data.

Prior: $\theta \sim U(0,1)$ ← f may also be considered a "prior"

Observe data $x_1, \ldots, x_n$ from $f(x;\theta)$

Update using Bayes Thm:

$$f(\theta|\vec{x}) = \frac{f(\vec{x}|\theta)\, f(\theta)}{f(\vec{x})}$$

($f(\theta)$: prior; $f(\vec{x})$: a number; $f(\vec{x}|\theta)$: this is the likelihood)

$$= \frac{\mathcal{L}(\theta)\, p(\theta)}{C} \propto \mathcal{L}(\theta)\, p(\theta)$$

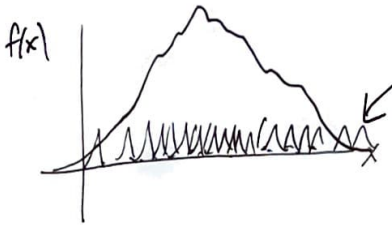
Often $f(\theta|\vec{x})$ is not a known distribution, and therefore sampling from it to compute, for example, an expectation (for a point estimate) can be difficult

⇒ the need for computation in statistics

→ mention hierarchical modelling

Density Estimation

Observe data $x_1, \ldots, x_n$, how do we estimate the underlying pdf that generated the data?



(1) How do you choose the bandwidth of the kernel? (e.g. cross-validation)

(2) What is a fast algorithm to evaluate $\hat{f}(x)$ at arbitary x?

(3) Extensions and algorithms for higher dimensions?

(4) How do we sample from $\hat{f}$?

Sampling Methods

Markov Chain Monte Carlo:

generate a sequence of random variables $X_1, X_2 \ldots$ (i.e. a Markov chain) such that as $n \to \infty$, $X_n \sim F$, we F is some specified distribution function. Not obvious how to do this in general.