Community Cheatsheet for 18.6501x

1 Algebra

Absolute Value Inequalities:
$|f(x)| < a \Rightarrow -a < f(x) < a$
$|f(x)| > a \Rightarrow f(x) > a$ or $f(x) < -a$

2 Calculus

2.1 Concavity in 1 dimension

If $g : I \to \mathbb{R}$ is twice differentiable in the interval I, i.e. $g''(x)$ exists for all $x \in I$, then g is

concave if and only if $g''(x) \leq 0$ for all $x \in I$;
strictly concave if $g''(x) < 0$ for all $x \in I$;
convex if and only if $g''(x) \geq 0$ for all $x \in I$;
strictly convex if $g''(x) > 0$ for all $x \in I$;

2.2 Multivariate Calculus

Gradient

Let

$$f : \mathbb{R}^d \longrightarrow \mathbb{R}\theta = \theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{pmatrix} \mapsto f(\theta)$$

denote a twice differentiable function, the Gradient ∇ of f is defined as:

$\nabla f : \mathbb{R}^d \to \mathbb{R}^d$

$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{pmatrix} \mapsto \begin{pmatrix} \frac{\partial f}{\partial \theta_1} \\ \frac{\partial f}{\partial \theta_2} \\ \vdots \\ \frac{\partial f}{\partial \theta_d} \end{pmatrix}\Bigg|_\theta$$

Hessian

The Hessian of f is the matrix $\mathbf{H} : \mathbb{R}^d \to \mathbb{R}^{d \times d}$ whose entry in the i-th row and j-th column is defined by

$$(\mathbf{H}f)_{ij} := \frac{\partial^2}{\partial\theta_i \partial\theta_j} f, \quad 1 \leq i, j \leq d$$

Semi-Definiteness

A symmetric (real-valued) $d \times d$ matrix $\mathbf{A}$ is:

Positive semi-definite if $\mathbf{x}^T \mathbf{A} \mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^d$.

Positive definite if inequality above is strict $\mathbf{x}^T \mathbf{A} \mathbf{x} > 0$ for all non-zero vectors $\mathbf{x} \in \mathbb{R}^d$

Negative semi-definite (resp. negative definite) if $\mathbf{x}^T \mathbf{A} \mathbf{x}$ is non-positive (resp. negative) for all $\mathbf{x} \in \mathbb{R}^d - \{0\}$.

Positive (or negative) definiteness implies positive (or negative) semi-definiteness.

Concavity

3 Important probability distributions

Bernoulli

Parameter $p \in [0, 1]$. Discrete, describes the success or failure in a single trial.

$$p_X(k) = \begin{cases} p, & \text{if k = 1} \\ (1-p), & \text{if k = 0} \end{cases}$$

$E[X] = p$
$Var(X) = p(1-p)$

Exponential

Parameter λ. Continuous

$$f_x(x) = \begin{cases} \lambda exp(-\lambda x), & \text{if x >= 0} \\ 0, & \text{o.w.} \end{cases}$$

$$F_x(x) = \begin{cases} 1 - exp(-\lambda x), & \text{if x >= 0} \\ 0, & \text{o.w.} \end{cases}$$

$E[X] = \frac{1}{\lambda}$
$Var(X) = \frac{1}{\lambda^2}$

Normal (Gaussian)

Parameters μ and $\sigma^2 > 0$. Continuous

$$f(x) = \frac{1}{\sqrt{(2\pi\sigma)}} exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

$E[X] = \mu$
$Var(X) = \sigma^2$
Useful properties:

Poisson

Parameter λ. Discrete, approximates the binomial PMF when n is large, p is small, and $\lambda = np$.

$(\mathbf{p_X}(k) = exp(-\lambda)\frac{\lambda^k}{k!}$ for $k = 0, 1, \ldots,$

$\mathbf{E}[X] = \lambda$
$Var(X) = \lambda$

Uniform

4 Expectation and Variance

Expectation

Variance

Covariance

Variance and expectation of mean of n iid random variables

Let $X_1, ..., X_n \overset{iid}{\sim} P_\mu$, where $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ for all $i = 1, 2, ..., n$ and $\overline{X_n} = \frac{1}{n}\sum_{i=1}^n X_i$.

Variance of the Mean:

$Var(\overline{X_n}) = (\frac{\sigma^2}{n})^2 Var(X_1 + X_2, ..., X_n) = \frac{\sigma^2}{n}$.

Expectation of the mean:

$E[\overline{X_n}] = \frac{1}{n}E[X_1 + X_2, ..., X_n] = \mu$.

5 LLN and CLT

Let $X_1, ..., X_n \overset{iid}{\sim} P_\mu$, where $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ for all $i = 1, 2, ..., n$

Weak and strong law of large numbers:

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow[n\to\infty]{P,a.s.} \mu .$$

$$\frac{1}{n}\sum_{i=1}^n g(X_i) \xrightarrow[n\to\infty]{P,a.s.} \mathbf{E}[g(X)]$$

Central Limit Theorem:

$$\sqrt{(n)}\frac{\overline{X_n} - \mu}{\sqrt{(\sigma^2)}} \xrightarrow[n\to\infty]{(d)} N(0,1)$$

$$\sqrt{(n)}(\overline{X_n} - \mu) \xrightarrow[n\to\infty]{(d)} N(0, \sigma^2)$$

6 Statistical models

7 Estimators

8 Confidence intervals

Onesided

Twosided

Delta Method

9 Hypothesis tests

Onesided

Twosided

P-Value

10 Distance between distributions

Total variation

The total variation distance TV between the propability measures P and Q with a sample space E is defined as:

$$\mathrm{TV}(\mathbf{P}, \mathbf{Q}) = \max_{A \subset E}|\mathbf{P}(A) - \mathbf{Q}(A)|,$$

Calculation with f and g:

$$\mathrm{TV}(\mathbf{P}, \mathbf{Q}) = \begin{cases} \frac{1}{2}\sum_{x\in E}|f(x) - g(x)|, & \text{discr} \\ \frac{1}{2}\int_{x\in E}|f(x) - g(x)|dx, & \text{cont} \end{cases}$$

Symmetry:
$d(\mathbf{P}, \mathbf{Q}) = d(\mathbf{Q}, \mathbf{P})$
nonnegative:
$d(\mathbf{P}, \mathbf{Q}) \geq 0$
definite:
$d(\mathbf{P}, \mathbf{Q}) = 0 \iff \mathbf{P} = \mathbf{Q}$
triangle inequality:
$d(\mathbf{P}, \mathbf{V}) \leq d(\mathbf{P}, \mathbf{Q}) + d(\mathbf{Q}, \mathbf{V})$
If the support of $\mathbf{P}$ and $\mathbf{Q}$ is disjoint:
$d(\mathbf{P}, \mathbf{V}) = 1$
TV between continuous and discrete r.v:
$d(\mathbf{P}, \mathbf{V}) = 1$

KL divergence

the KL divergence (also known as relative entropy) KL between between the propability measures P and Q with the common sample space E and pmf/pdf functions f and g is defined as:

$$\mathrm{KL}(\mathbf{P}, \mathbf{Q}) = \begin{cases} \sum_{x\in E} p(x)\ln\left(\frac{p(x)}{q(x)}\right), & \text{discr} \\ \int_{x\in E} p(x)\ln\left(\frac{p(x)}{q(x)}\right)dx, & \text{cont} \end{cases}$$

Not a distance!
Asymetric in general:
$\mathrm{KL}(\mathbf{P}, \mathbf{Q}) \neq \mathrm{KL}(\mathbf{Q}, \mathbf{P})$
Nonnegative:
$\mathrm{KL}(\mathbf{P}, \mathbf{Q}) \geq 0$
Definite:
if $\mathbf{P} = \mathbf{Q}$ then $\mathrm{KL}(\mathbf{P}, \mathbf{Q}) = 0$
Does not satisfy triangle inequality in general:
$KL(\mathbf{P}, \mathbf{V}) \nleq KL(\mathbf{P}, \mathbf{Q}) + KL(\mathbf{Q}, \mathbf{V})$

Estimator of KL divergence:
$KL(\mathbf{P}_{\theta_*}, \mathbf{P}_\theta) = const - \mathbf{E}[ln(p_\theta(X))]$

$\widehat{KL}(\mathbf{P}_{\theta_*}, \mathbf{P}_\theta) = const - \frac{1}{n}\sum_{i=1}^n log(p_\theta(X_i))$

11 Likelihood

Let $(E, \{P_\theta\}_{\theta\in\Theta})$ denote a discrete or continuous statistical model. Let p_θ denote the pmf or pdf of P_θ. Let $X_1, \ldots, X_n \overset{iid}{\sim} P_{\theta^*}$ where the parameter θ^* is unknown. Then the likelihood is the function

$L_n : E^n \times \Theta$
$(x_1, \ldots, x_n, \theta)$
$L_n(x_1, \ldots, x_n, \theta) = \prod_{i=1}^n P_\theta[X_i = x_i]$

Discrete Likelihood

Bernoulli likelihood:
$L_n(x_1, \ldots, x_n, p) = p^{\sum_{i=1}^n x_i}(1-p)^{n - \sum_{i=1}^n x_i}$
$= L_n = \prod_{i=1}^n (x_i p + (1 - x_i)(1-p))$
Poisson likelihood:
$L_n(x_1, \ldots, x_n, \lambda) = \prod_{i=1}^n \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} e^{n\lambda}$

Poisson loglikelihood:
$log(L(x_1 \ldots x_n; \lambda)) =$
$-n\lambda + log(\lambda)(\sum_{i=1}^n x_i)) - log(\prod_{i=1}^n x_i!)$

Continuous Likelihood

Gaussian likelihood:

$$L(x_1 \ldots x_n; \mu, \sigma^2) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right)$$

Gaussian loglikelihood:

$log(L(x_1 \ldots x_n; \mu, \sigma^2)) =$
$-nlog(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2$

Gaussian Maximum-loglikelihood estimators:

MLE estimator for $\sigma^2 = \tau$:
$\hat{\tau}_n^{MLE} = \frac{1}{n}\sum_{i=1}^n X_i^2$

MLE estimators:

$\hat{\mu}_n^{MLE} = \frac{1}{n}\sum_{i=1}(x_i)$
Exponential likelihood:
$L(x_1 \ldots x_n; \lambda) = \lambda^n \exp\left(-\lambda\sum_{i=1}^n x_i\right)$
Uniform:
$L(x_1 \ldots x_n; b) = \frac{1(\max_i(x_i \leq b))}{b^n}$