Week 1 - 20/04/2020

- Youtube 31-08-2016
- Conclude week 1 L1 - review, begin (W1)(L2) - inequalities
-

(1) $A_z = \{(x,y)\ g(x,y) \overset{(=?)}{\leq} z\}$

- Notation:-

$X \perp Y$

(2)
- Independence can be physical fact OR an assumption which we have to eval.
- How to check for independence?
- Joint density factors multiplicatively $p_{X,Y}(x,y) = p_X(x)\, p_Y(y)$
- Easier to check independence against density/distribution than in terms of events
- Same for multiple iid X_is.
- Identical distri - n observations drawn from same population / same marginal distribution
- iid - independent, identically distributed; random sample from population / common distribution
- An assumption

(3)
- IID notation:

$X_1, \ldots, X_n \sim P$ (iid drawn from common probab. distribution P) or $X_1, \ldots, X_n \sim F$ (iid drawn from same CDF) or $X_1, \ldots, X_n \sim p$ (iid from (prob) same density function)

(2). By default in the class - assume iid

- Non iid: departing from this is very important e.g. time series

(4): Distributions included for reference

- Normal distribution is a family of distn indexed by 2 parameters μ, σ

- $p(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ $\quad \mathbb{E}[X] = \mu$

x indicates r.v.; μ, σ indicates parameter $\quad var(X) = \sigma^2$

MVN: Radom vectors

- vector of radom variables $\underline{X} = \begin{pmatrix} X(1) \\ \vdots \\ X(d) \end{pmatrix} \in \mathbb{R}^d$

- Typo ⑤: $p(\underline{x}) = \frac{1}{(2\pi)^{d/2} |\underline{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\underline{x}-\underline{\mu})^T \underline{\Sigma}^{-1} (\underline{x}-\underline{\mu})\right)$

- $\underline{\mu} = \begin{pmatrix} \mu(1) \\ \vdots \\ \mu(d) \end{pmatrix}$ $\underline{\Sigma}$ = variance covariace matrix $= \begin{bmatrix} \quad \end{bmatrix}$

- very important facts:- MVN (review) cd prop.
- marginal distn of any set of coordinates is Normal
- condit. distn of any —"— given set of coordinates is also Normal

⑥ χ^2 distn

- $z_1, \dots, z_p \sim N(0,1)$ - χ^2 distn is a family of distn indexed by parameter p

$X = \sum_{j=1}^{p} z_j^2$

then $X \sim \chi^2_p$

$\mathbb{E}[\chi^2_p] = p$

$\mathrm{Var}[\chi^2_p] = 2p$

- Above a continuous distn.

onto discrete

Bernoulli $X \sim \mathrm{Bern}(\theta)$ if $P(X=1) = \theta$ $P(X=0) = 1-\theta$

$p(x;\theta) = \theta^x (1-\theta)^{1-x}$ $x = 0,1$

Binomial $X \sim \mathrm{Binomial}(\theta)$ (sum of n Bernoullis)

$p(x;\theta) = P(X=x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$ $x \in \{0, \dots, n\}$

- Remind yourself of these!

Sampling distributions

(STT)

- $X_1, \dots, X_n \sim P$

"Take a sample of data, take function of sample, which itself defines new r.v." (summarise data)

- sample mean:

- A statistic can be viewed as a function/transform of data

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = g(X_1, \dots, X_n)$$

- As $X_1, \dots, X_n$ are random; $\bar{X}_n$ is also random

· $\bar{X}_n$, as an r.v.; also has a distri, the sampling distri

$G_n(t) = P(\bar{X}_n \leq t)$

· In general, sampling distri is difficult to find, in particular as P is unknown

· However sampling distri has properties:-

- See practice problem for completion

23:33

(Q1): Difference between distribution of the X_is and the $\bar{X}_n$s

· sample variance: $S_n = \frac{1}{n-1}\sum (X_i - \bar{X})^2 \Rightarrow \mathbb{E}(S^2) = \sigma^2$

· why divide by $(n-1)$ instead of $n \rightarrow$ unbiased

· only assuming ~~the~~ data iid, mean and variance exist

$\mathbb{E}(S^2) = \sigma^2$

· Theorem 10 builds on this by specifying that $X_1, \dots, X_n \sim N(\mu, \sigma^2)$

(a) we can say ~~$\bar{X}$ has~~ $\bar{X}_n \sim N(\mu, \sigma^2/n)$ (STT)

· Proved by mgf of $\bar{X}_n$; mgfs completely characterise a distribution

· That is, compute mgf $\Rightarrow \bar{X}_n \sim N(\mu, \sigma^2/n)$

(b): $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$

- contents of Theorem 10 - only if $X_1, \dots, X_n \sim N(\mu, \sigma^2)$

(Q): 28:20 $\rightarrow$ query about intersections; and clarity on double integral

Week 1 - L2 Youtube 31-08-2016 20/04/202

- Probability Inequalities

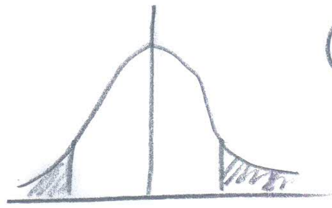
- Finding bounds
- Probability bounds useful in ML ~~in particular~~ Hoeffding
- Stepping stone to VC Theory (advanced prob. bounds) → convergence theory (what happens to $\bar{X}_n$ when more data)

probability bounds →

- Multitude of bounds; but main:-
  1) Markov
  2) Hoeffding
  3) Gaussian

31:50?

Theorem 1 $X \sim N(0,1)$



due to $|X_n|$

OT
- Tails
- How fast do the tails
- LW uses t (arbitrary positive real t)

- As $t\uparrow$, $P(|X_n|>t)$ gets small (very important)

- 2nd part: intuitively:
  $P(|\bar{X}_n|>t)$ goes to 0 quickly in terms of t and sample size n

(LW) inequalities tend to mimick Gaussian tail inequalities for other distributions

Theorem 1 (Proof)

- Do one side, use symmetry → factor 2

(LW): $\phi(s)$ is stand normal density

- multiply/divide by s
- check 36:11 latter step

Property: OT: $\phi'(s) = -s\phi(s)$

$\Rightarrow s\phi(s) = -\phi'(s)$

(A1)
(Q): check proof that you understand steps ✓

Theorem 2 - Markov's inequality

- very weak, but used to prove Hoeffding $\mathbb{E}(X)$
- Assume $X > 0$ and that mean exists and is well defined i.e. finite mean
- (W): Cauchy distribution - tails spread out very slowly (A2) (WIKI)
  - does not have well-defined mean due to fat-tails
- As $X > 0$, integrate in range $[0, \infty)$

(A3) check proof

- As $t \uparrow$; $P(X>t) \downarrow$ by $\frac{1}{t}$ → quantify weakness

Theorem 3 - Chebyshev's inequality

- Assuming that variance is well defined and exists (i.e. finite)
  - this extra information can be used
- Chebyshev's inequality builds/augments Markov inequality using this info to get a sharper inequality

Hoeffding inequality

- sharper; assume that X is bounded above and below
- Boundedness ⇒ thin tails
- Probability mass is 0 outside interval $[a,b]$; tails cannot be thinner
- improvement over Markov, Chebyshev $P(|\bar{X}_n - \mu| \geq \epsilon) \leq 2e^{-2n\epsilon^2/(b-a)^2}$
- Proof is useful (useful tricks)
- 2 tricks: 1. All moments of X exists (moment generating function exist?)
  2. How to bound mgf using bound on X
- Chernoff's method (going back to epsilon)

$$P(X > \epsilon) \leq \inf_{t \geq 0} (\text{"min"})\ e^{-t\epsilon}\ \mathbb{E}[e^{tX}]$$

mgf

$t \geq 0$ variational param.

- both are functions of t
- minimise over t

- check 50:06 for chernoff

14 - check you undestand chernoff before next lecture