LAST WEEK

Two result (No proof)

1) LLN $\quad S_N = \frac{1}{N}\sum_i X_i \xrightarrow{P} \mathbb{E}[X]$

$$\mathbb{P}\left[\left| S_N - \mathbb{E}[X]\right| > \varepsilon\right] \to 0$$

2) CLT $\quad \frac{S_N - N}{\sqrt{\Delta}}\sqrt{N} \xrightarrow{D} Z \sim \mathcal{N}(0,1)$

Comparing Average : how far S_N is from $\mu = \mathbb{E}[X]$?

In General : $\mathbb{P}[|S_N - \mu| \geq k\sqrt{\Delta}] \leq \frac{1}{k^2}$ (Chebyshev)

Can be MUCH Better ; Ex X Gaussian $\sim \mathcal{N}, \Delta$

$$\mathbb{P}\left[\frac{1}{N}\sum X_i > \mu + \varepsilon\right] = \mathbb{P}\left[\sum_i X_i \geq N(\mu+\varepsilon)\right] = \mathbb{P}\left[e^{t\sum_i X_i} \geq e^{Nt(\mu+\varepsilon)}\right] \quad t \geq 0$$

$$\leq \frac{\mathbb{E}\left[e^{t\sum_i X_i}\right]}{e^{Nt[\mu+\varepsilon]}} = \left(\mathbb{E}\left[e^{tX}\right] e^{-t(\mu+\varepsilon)}\right)^n \quad \text{(Markov)}$$

$$= \left(e^{t\mu + \frac{t^2}{2\Delta}} e^{-t(\mu+\varepsilon)}\right)^n = e^{n\left[\frac{t^2}{2\Delta} - t\varepsilon\right]} \quad \text{(MGF)}$$

$\frac{t^*}{\Delta} = \varepsilon$

$$\mathbb{P}\left[\frac{1}{N}\sum_i X_i \geq \mu + \varepsilon\right] \leq e^{-\frac{n\varepsilon^2}{2\Delta}}$$

Similarly $\quad \mathbb{P}\left[\frac{1}{N}\sum_i X_i \leq \mu - \varepsilon\right] \leq e^{-\frac{n\varepsilon^2}{2\Delta}}$

$$\mathbb{P}\left[\left|\frac{1}{N}\sum_i X_i - \mu\right| \geq \varepsilon\right] \leq 2e^{-\frac{n\varepsilon^2}{2\Delta}}$$

$$\mathbb{P}\left[\left|\frac{1}{N}\sum_i X_i - \mu\right| \leq \varepsilon\right] \geq 1 - \underbrace{2e^{-\frac{n\varepsilon^2}{2\Delta}}}_{\delta}$$

Accuracy is cheap

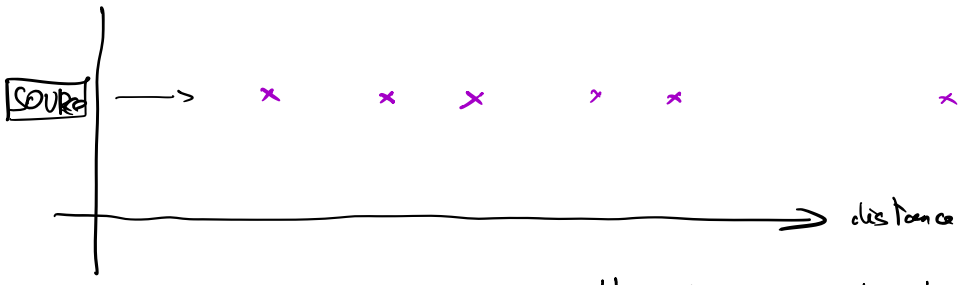
$\varepsilon = \sqrt{\frac{2\Delta}{n}\log\frac{2}{\delta}}$

with prob at least $1-\delta$, $|\hat{p}-p| \le \sqrt{\frac{2\Delta}{m}\log\frac{2}{\delta}}$ ← Precision is Expensive

PAC Learning

Lecture II All of Statistics

Learn a single parameter



How do you estimate the "mean-path" typical distance θ^*

$$\hat{\theta}_{EM}(\{x_i\}) = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$P(x;\theta^*) = \frac{1}{\theta^*} e^{-\frac{x}{\theta^*}}$$ mean and variance

$$\mathbb{E}\left[\hat{\theta}_{EM}(\{x_i\}_{i=1}^{m=6})\right] = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}x_i = \theta^*$$

$$\text{VAR}\left[\hat{\theta}^m_{EM}\right] = \frac{1}{m}\text{VARIANCE}(x) = \frac{\theta^{*2}}{m}$$

$$\hat{\theta}^m_{ML} = \theta^* \pm \frac{\theta^*}{\sqrt{m}}$$

How to choose $\hat{\theta}(\{x_i\})$??

1 Bayesian Approach

likelihood

Prior

$$P(\theta|\{x_i\}) = \frac{P(\{x_i\}|\theta)\,P(\theta)}{P(\{x_i\})} \quad \text{(Z)}$$

Posterior Distribution

$$\tilde{P}(x|\theta) = \begin{cases} x>20 \\ x<1 \end{cases} 0 \qquad 20>x>1 \quad \frac{1}{Z(\theta)}\frac{1}{\theta}e^{-x/\theta} = \frac{e^{-x/\theta}}{\int_1^{20} dx\, e^{-x/\theta}}$$

$$P(\{x_i\}|\theta) = \prod_{i=1}^{n} \tilde{P}(x_i|\theta)$$

Jeffrey Prior ?

$P(\theta) = \text{cste}$ — "Maximum entropy" / "minimum Assumption"

???

$P(\theta|\{x_i\}) \propto \text{likelihood} \Rightarrow$

$$\hat{\theta}_{ML} = \text{ARGMAX}\, P(\{x_i\}|\theta)$$

Maximum likelihood

$$\hat{\theta}_{ML} = \text{ARGMAX}\, \log P(\{x_i\}|\theta)$$

(2) Frequentist

$$\text{Risk}\,[\hat{\theta}, \theta^*] = \mathbb{E}\left(\hat{\theta}(\{x\}) - \theta^*\right)^2 \quad \text{QUADRATIC RISK}$$

MSE
Mean Squared Error

Study of maximum likelihood

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}} \log \prod_{i=1}^{n} P(x_i \mid \theta)$$

$$\ell(x, \theta) = \log P(x \mid \theta)$$
$$d^n(\{x\}, \theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, \theta)$$

$$\hat{\theta}^n(\{x\}) = \text{ARGMAX}\; d_n(\{x\}, \theta)$$

1/ Consistent $n \to \infty$, then $\hat{\theta}_{ML} \longrightarrow \theta^*$

$$\lim_{n\to\infty} d^n(\{x\}, \theta) - d^n(\{x\}, \theta^*)$$

$$\lim_{n\to\infty} = \frac{1}{n}\sum_{i=1}^{n} \log P(x_i \mid \theta) - \frac{1}{n}\sum_{i=1}^{n} \log P(x_i \mid \theta^*)$$

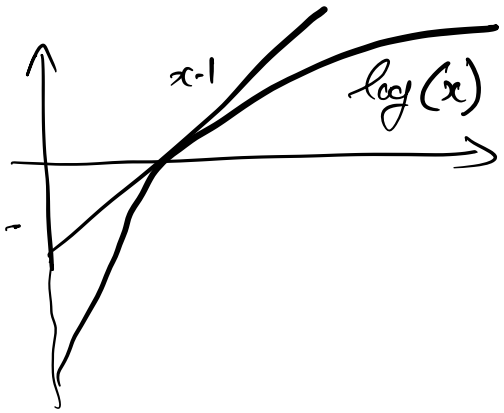
$$= \mathbb{E} \log P(x \mid \theta) - \mathbb{E} \log P(x \mid \theta^*)$$

$$= \int dx\, P(x \mid \theta^*) \log P(x \mid \theta) - \int dx\, P(x \mid \theta^*) \log P(x \mid \theta^*)$$

$$= \int dx\, P(x \mid \theta^*) \log \frac{P(x \mid \theta)}{P(x \mid \theta^*)} = -D_{KL}\left(P(x \mid \theta^*) \| P(x \mid \theta)\right)$$

$$\leq 0$$

Proof Gibbs Inequality



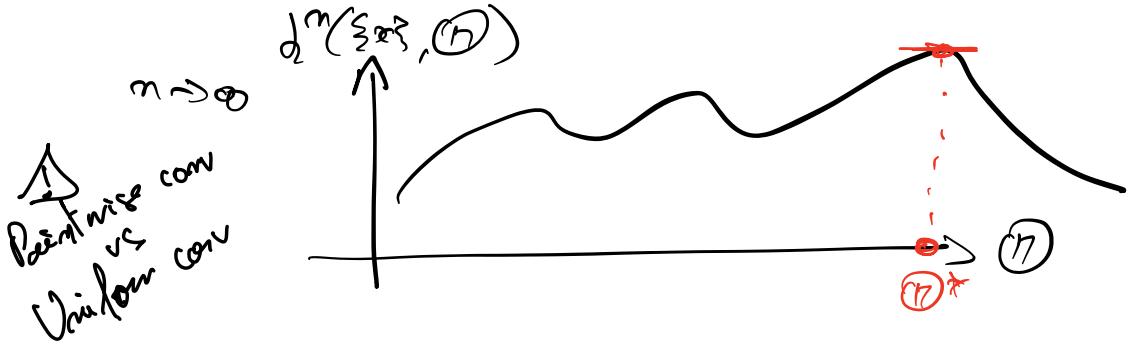
$x-1 \geq \log x$

$p_i , q_i \qquad i = 1 \dots N$

$x = \frac{p_i}{q_i}$

$$\sum_{i=1}^{N} q_i \left(\frac{p_i}{q_i} - 1 \right) \geq \sum_{i=1}^{N} q_i \log \frac{p_i}{q_i}$$

$$0 = \sum_{i=1}^{N} p_i - \sum_{i=1}^{N} q_i \geq \sum_{i=1}^{N} q_i \log \frac{p_i}{q_i}$$



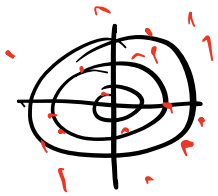
MAXIMUM LIKELIHOOD IS CONSISTENT

$$\lim_{n \to \infty} \Theta_{ML}^{n}(\xi x^n) = \Theta^*$$

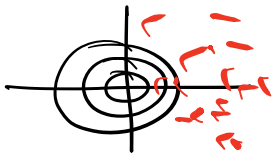
IV Limits of learning

BIAS-VARIAN

$$MSE = E\left[\left(\hat{\lambda}(\{x_i\}) - \lambda^*\right)^2\right] = \underbrace{E\left[\hat{\lambda}(\{x_i\})^2\right] - E\left[\hat{\lambda}\{x_i\}\right]^2}_{\text{VARIANCE}(\hat{\lambda})} + \underbrace{\left(E\left[\hat{\lambda}(\{x_i\})\right] - \lambda^*\right)^2}_{\text{BIAS}}$$

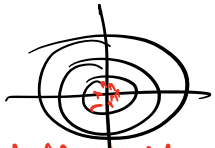


large variance
small bias



large variance
large bias

small variance
large bias



small variance, small bias

$$\text{bias}(\hat{\theta}, \theta) = E(\hat{\theta} - \theta^*)$$

$$MSE(\theta^*, \hat{\theta}) \geq b^2(\hat{\theta}, \theta^*) + \frac{1}{nI(\theta^*)}(1 + \gamma_\theta \text{bias})^2$$

CRAMERS - RAO

Assume n is large, and assume $\hat{\lambda}$ is consistent when $n \to \infty$

$\hat{\lambda}$ should be unbiased

Bound for unbiased Estimators

$$MSE \geq \frac{1}{nI(\theta^*)}$$

Fischer Score

$$S(x,\theta) = \frac{\partial \ell(x,\theta)}{\partial \theta} \qquad S^n(\{x\},\theta) = \frac{\partial \ln(\{x\},\theta)}{\partial \theta}$$

$$\mathbb{E}[S(x,\theta^*)] = 0$$

$$\mathbb{E}[S^2(x,\theta^*)] = -\mathbb{E}[\partial_\theta S(x,\theta)|_{\theta^*}] = I_{Fi}(\theta^*) \quad \text{Fisher Information}$$

$$\mathbb{E}\left[\left(S^n(\{x\},\theta^*)\right)^2\right] = n\, I_{Fi}(\theta^*)$$

$$\int dx\, P(x|\theta^*)\, \frac{\partial_\theta P(x|\theta)|_{\theta^*}}{P(x|\theta^*)}$$

$$= \partial_\theta \underbrace{\int dx\, P(x|\theta^*)}_{=1} = 0$$

$$\mathbb{E}[\partial_\theta S(x,\theta)|_{\theta^*}] = \int dx\, P(x|\theta^*)\, \frac{(\partial_\theta P)P - (\partial_\theta P)^2}{P^2(x|\theta^*)}$$

$$= \int dx\, \partial_\theta^2 P - \int dx\, P\left(\frac{\partial_\theta P}{P}\right)^2 = -\mathbb{E}[S^2]$$

Proof CR : Cauchy-Swartz

$$\mathrm{COV}[\hat{\theta}, S]^2 \leq \mathrm{VAR}[\hat{\theta}]\,\mathrm{VAR}[S] = \mathrm{VAR}(\hat{\theta})\, n I$$

$$\mathrm{VAR}[\hat{\theta}] \geq \frac{(\mathbb{E}[\hat{\theta} S] - \mathbb{E}[\hat{\theta}]\mathbb{E}[S])^2}{n I}$$

$$MSE \geq \frac{\mathbb{E}[\hat{\theta} S]^2}{n I} + bias^2$$

But $\Rightarrow$ $\partial\, bias = \partial_\theta \mathbb{E}[\hat{\theta}] \cdot 1 = \partial_\theta \int dx\, P(x|\theta)\hat{\theta}(x) \cdot 1 = \int dx\, P' \hat{\theta} \cdot 1$

$$= \int dx\, P \frac{P'}{P} \hat{\theta} \cdot 1 = \mathbb{E}[S\hat{\theta}] - 1 \qquad \blacksquare$$

3/ M.L is Efficient => It Achieves Cramer-Rao as $n \to \infty$
Asymptotically

Bound for Unbiased Estimators $\quad MSE \geq \frac{1}{nI(\theta^*)}$

$$0 = \dot{\lambda}^n(\{x_i\}, \hat{\theta}_{ML}) = \dot{\lambda}^n(\{x_i\}, \theta^*) + (\hat{\theta}_{ML} - \theta^*)\ddot{\lambda}(\{x_i\}, \theta^*) + O(\theta - \theta)^2$$

$$\sqrt{n}(\hat{\theta}_{ML} - \theta^*) = -\sqrt{n}\frac{\dot{\lambda}_n(\{x_i\}, \theta^*)}{\ddot{\lambda}_n(\{x_i\}, \theta^*)} + \times$$

CLT $\rightarrow$

$$\sqrt{n}(\hat{\theta}_{ML} - \theta^*) = -\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{\ell}(x_i, \theta^*)}{\frac{1}{n}\sum_{i=1}^{n}\ddot{\ell}(x_i, \theta^*)} \qquad \mathcal{N}(0, I)$$

$\rightarrow I(\theta^*)$

$$\sqrt{n}(\hat{\theta}_{ML} - \theta^*) \xrightarrow[n \text{ large}]{} \mathcal{N}(0, \frac{1}{I})$$

$\hat{\theta}_{ML}$ distributed as $e^{-\frac{(\hat{\theta}_{ML} - \theta)^2}{2\frac{1}{nI}}} / \sqrt{2\pi\frac{1}{nI}}$

$$\hat{\theta}_{ML} \sim \mathcal{N}(\theta^*, \frac{1}{nI})$$

① Maximum likelihood is Asymptotically Gaussian
② " " achieves " " Cramer-Rao Bound

ML is efficient

Final Note $\Rightarrow$ Work for Many Variables as Well

$$\left[I(\vec{\theta})\right]_{ij} = E\left[\frac{\partial \log P}{\partial \theta_i}\frac{\partial \log P}{\partial \theta_j}\right] = -E\left[\frac{\partial}{\partial\theta_i\partial\theta_j}\log P\right]$$

Fisher Matrix

$$\mathrm{Cov}[\hat{\theta}] \geq (I_{ij})^{-1}$$ ← Cramer-Rao