Today

1. How we’re going to do things
2. Universal Coding/Compression (powerpoint)
   – the normalized maximum likelihood code for finite/countable models

   (Chapter 6, Section 6.1-6.2)
3. Kullback-Leibler divergence,Fisher information and squared Euclidean distance (Chapter 4, 4.1-4.3) (written on paper)
4. Questions/Feedback

Something about Epidemiology

- Epidemics involve exponential growth
- R_0 : Basic reproduction nr
  - nr of persons that an infected person infects, on average
- Corona R_0 : about 2.5 (maybe even 4)
- Social Distancing: getting R_0 down below one

- Major problem: are the measures taken good enough?
- Phase transitions:
  - 70% social distancing may imply ‘IC (intensive care) wards don’t fill up’
  - whereas 75% social distancing may imply ‘2000 people in NL need to but cannot go to IC’ (this means they’ll die)
- …if R_0 is 2.4. If $R_0 = 3$ it is completely different again.

- Major problem: are measures taken good enough?
- Second major problem: effect of measures only visible after 10-14 days. In mean time, exponential growth may continue
- We are living in complete uncertainty about this right now.
  - Situation (how many elderly, how many hospital beds, effect of social distancing) very different from country to country
  - Therefore harsh measures are inevitable

Corona is not at all like the flu!

- If everybody gets the attention they need, then like with the flu, it mostly kills very old people that were already ill
- Because of higher R_0 , much longer incubation time and much higher percentage of people that need to go to hospital, there is a substantial chance that many (also young) people don't get the attention they need. And then they die.

Universal Codes

- $\mathcal{L}$: set of code (length function)s available to encode data $x^n = (x_1, \ldots, x_n)$
- Suppose we think that one of the code(length function)s in $\mathcal{L}$ allows for substantial compression of x^n
- GOAL (for now): encode x^n using minimum number of bits!

Universal **Codes**

- Simply encoding x^n using the $\hat{L} \in \mathcal{L}$
  that minimizes code length
  does not work (encoding cannot be decoded)

- But there exist codes L which, for any sequence x^n are 'almost' as good as $\inf_{L \in \mathcal{L}} L(x^n)$

- These are called **universal codes** for $\mathcal{L}$

Universal Codes

- Example: $\mathcal{L}$ finite
- There exists 2-part code $L_{2-\mathrm{p}}$ such that for some constant K, all n , all $x^n \in \mathcal{X}^n$

$$L_{2\text{-p}}(x^n) \leq L(x^n) + K$$

- In particular,

$$L_{2\text{-p}}(x^n) \leq \inf_{L \in \mathcal{L}} L(x^n) + K$$

- IMPORTANT: K does not depend on n , while typically, for all n, $L(x^n)$ grows linearly in n

Universal Models

- Let $\mathcal{M} = \{ p_\theta : \theta \in \Theta\}$ be a probabilistic model, i.e. a family (set) of probability distributions
- Assume $\mathcal{M}$ finite
- By Kraft inequality applied to p_θ, there exists code $L_{2-\mathrm{p}}$ such that for all n , all $x^n \in \mathcal{X}^n$

$$L_{\text{2-p}}(x^n) \leq \inf_{\theta \in \Theta} - \log p_\theta(x^n) + K$$

- Hence exists a (defective) distribution such that

$$- \log p_{\text{2-p}}(x^n) \leq \inf_{\theta \in \Theta} - \log p_\theta(x^n) + K$$

i.e. $\quad p_{\text{2-p}}(x^n) \geq K' \cdot p_{\hat{\theta}(x^n)}(x^n)$

Terminology

- Statistics (and in my book):
  - Model = family of distributions
- Information theory:
  - Model = single distribution
  - Model class = family of distributions
- So in my book: universal model is a single distribution acting as a representative of/defined relative to a set of distributions

Bayesian Mixtures are universal models

- Let w be a prior over Θ . The Bayesian marginal likelihood is defined as:

$$\bar{p}_{\text{Bayes}}(x^n) = \sum_{\theta \in \Theta} p_\theta(x_1, \ldots, x_n) w(\theta)$$

- This is a universal model, since for all $\theta_0 \in \Theta$:

$$- \log \bar{p}_{\text{Bayes}}(x^n)$$
$$= - \log \sum_{\theta} p_\theta(x^n) w(\theta) \leq - \log p_{\theta_0}(x^n) - \log w(\theta_0)$$

so

$$- \log \bar{p}_{\text{Bayes}}(x^n)$$
$$\leq \inf_{\theta \in \Theta} \{- \log p_\theta(x^n) - \log w(\theta)\} \leq - \log p_{\hat{\theta}(x^n)}(x^n) - \log w(\hat{\theta}(x^n))$$

2-part MDL code is universal also with nonuniform code on Θ

- Code x^n by first coding $\hat{\theta}(x^n)$, the maximum likelihood estimate, then code 'with the help of' $\hat{\theta}(x^n)$:

$$\begin{aligned} L_{2\text{-p}}(x^n) &:= \inf_{\theta \in \Theta} -\log p_\theta(x^n) - \log w(\theta) \\ &\leq -\log p_{\hat{\theta}(x^n)}(x^n) - \log w(\hat{\theta}(x^n)). \end{aligned}$$

2-part vs. Bayes universal models

- Bayes' mixture typically 'better' universal model in that it assigns larger probability (shorter code length) to outcomes.

$$\text{Bayes: } -\log \sum_{\theta} p_\theta(x^n) w(\theta) < \inf_{\theta \in \Theta} \{ -\log p_\theta(x^n) \ - \log w(\theta) \}$$

$$\text{two-part} = \inf_{\theta \in \Theta} \{ -\log p_\theta(x^n) - \log w(\theta) \ \}$$

- But what does 'better' really mean?
- What *prior* leads to short code lengths?

Optimal Universal Model

Look for $\bar{p}$ such that worst-case regret

$$\sup_{x^n \in \mathcal{X}^n} \left\{ -\log \bar{p}(x^n) - (-\log p_{\widehat{\theta}(x^n)}(x^n)) \right\}$$

is small *no matter what* x^n *are;* i.e. look for

$$\inf_{\bar{p}} \sup_{x^n \in \mathcal{X}^n} \left\{ -\log \bar{p}(x^n) - (-\log p_{\widehat{\theta}(x^n)}(x^n)) \right\}$$

Optimal Universal Model - II

$$\inf_{\bar{p}} \sup_{x^n \in \mathcal{X}^n} \left\{ -\log \bar{p}(x^n) - (-\log p_{\hat{\theta}(x^n)}(x^n)) \right\}$$

is achieved for Normalized Maximum Likelihood (NML) distribution (Shtarkov 1987):

$$\bar{p}_{\text{nml}}(x^n) = \frac{p_{\hat{\theta}(x^n)}(x^n)}{\sum_{y^n \in \mathcal{X}^n} p_{\hat{\theta}(y^n)}(y^n)}$$

Optimal Universal Model - II

$$\inf_{\bar{p}} \sup_{x^n \in \mathcal{X}^n} \left\{ -\log \bar{p}(x^n) - (-\log p_{\widehat{\theta}(x^n)}(x^n)) \right\}$$

is achieved for Normalized Maximum Likelihood (NML) distribution (Shtarkov 1987):

$$\bar{p}_{\text{nml}}(x^n) = \frac{p_{\widehat{\theta}(x^n)}(x^n)}{\sum_{y^n \in \mathcal{X}^n} p_{\widehat{\theta}(y^n)}(y^n)}$$

For all x^n, regret given by

$$-\log \bar{p}_{\text{nml}}(x^n) - [-\log p_{\widehat{\theta}(x^n)}(x^n)] = \log \sum_{y^n \in \mathcal{X}^n} p_{\widehat{\theta}(y^n)}(y^n)$$

(equalizer strategy)

How do the three Universal Codes Compare for finite model, $|\Theta| = K$?

- 2-part: worst-case regret bounded by $\log K$
- Bayes: worst-case regret (usually strictly) smaller

- NML: worst-case regret given by parametric complexity

$$\mathrm{comp}(\mathcal{M}) = \log \sum_{y^n \in \mathcal{X}^n} p_{\hat{\theta}(y^n)}(y^n)$$

- even (usually strictly) smaller

See Sheet

- 2-part code "syntactic"
- NML code "semantic": if all distributions are 'close' distributions, $\ll \log|\Theta|$
- Next week (and already in homework): NML with infinite Θ. NML idea still works
- For 'parametric' models with 'compact' Θ ,

$$\text{comp}(\mathcal{M}) = \log \sum_{y^n \in \mathcal{X}^n} p_{\hat{\theta}(y^n)}(y^n)$$

  typically grows with n (logarithmically)
- We have seen 4 types of universal code ; in 2 weeks a 5th one.

$$\sum_{x^n \in \mathcal{X}^n} P_{\hat{\Theta}(x^n)}(x^n)$$

$$= \sum_{\theta \in \Theta} \sum_{x^n : \hat{\Theta}(x^n) = \theta} P_\theta(x^n)$$

$$= \sum_{\theta \in \Theta} P_\theta(\hat{\Theta}(x^n) = \theta)$$

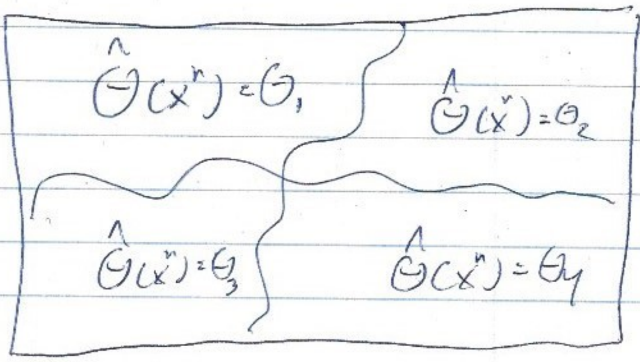
$$= \sum_{\theta \in \Theta} \left(1 - P_\theta(\hat{\Theta}(x^n) \neq \theta)\right)$$

$$= |\Theta| - \underbrace{\sum_{\theta \in \Theta} P_\theta(\hat{\theta}(x^n) \neq \theta)}_{\text{TOTAL AMOUNT OF CONFUSION}}$$

$\leq e^{-nC}$

EG

$\Theta = \{0.2, 0.4, 0.6, 0.8\}$



FISHER INFORMATION

$\mathcal{M} = \{ P_\Theta : \Theta \in \Theta \}$

$$I(\overset{\circ}{\Theta}) := \underset{Z \sim P_{\overset{\circ}{\Theta}}}{E} \left[\frac{d^2}{d\Theta^2} \left(-\log P_\Theta(Z) \right) \right]_{\Theta = \overset{\circ}{\Theta}}$$

BERNOULLI: $I(\Theta) = \frac{1}{\Theta(1-\Theta)} \left(= \frac{1}{\text{VARIANCE}_\Theta} \right)$

NORMAL LOCATION FAMILY : $I(\Theta) = \text{CONSTANT} \left(= \frac{1}{\text{VARIANCE}_\Theta} \right)$

$$D(\Theta_0 \| \Theta') = \underset{Z \sim P_{\Theta_0}}{E} \left[-\log P_{\Theta'}(Z) + \log P_{\Theta_0}(Z) \right]$$

$$= \underset{\Theta_0}{E} \Big[-\log P_{\Theta_0} + (\Theta' - \Theta_0) \frac{d}{d\Theta} \left[-\log P_\Theta(Z) \right]_{\Theta = \Theta_0}$$

$$+ \frac{1}{2} (\Theta' - \Theta_0)^2 \frac{d^2}{d\Theta^2} \left[-\log P_\Theta(Z) \right]_{\Theta = \Theta_0} + \langle \text{SMALL} \rangle$$

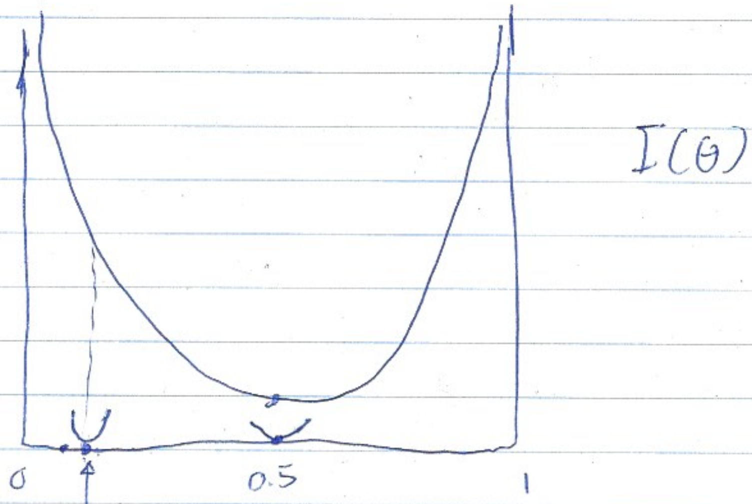
$$+ \underset{\Theta_0}{E} \left[\log P_{\Theta_0}(Z) \right]$$

$$= \frac{1}{2} I(\Theta_0) (\Theta_0 - \Theta')^2 + \text{REST}.$$

BOOK: ALSO MULTIDIMENSIONAL

$$D(\theta_0 \| \theta) \approx \frac{1}{2} I(\theta)(\theta_0 - \theta)^2$$

EXAMPLE BERNOULLI



$\theta = 0.5: D(\theta \| \theta + \varepsilon) \approx \frac{1}{4}\varepsilon^2$.

$\theta = 0.1: D(\theta \| \theta + \varepsilon) \approx \frac{10}{0.9}\varepsilon^2 \approx 11\varepsilon^2$

NORMAL: $I(\theta)$ CONSTANT

LOCATION

$$D(\theta \| \theta') = \frac{1}{2\sigma^2}(\theta - \theta')^2$$

"NO CURVATURE"

NEXT WEEK

- TWO PART DISCRETIZE WITH WIDTH $\propto \sqrt{I(\theta)}$ $\sqrt{\det I(\theta)}$
- BAYES: YOU USE PRIOR $\propto \sqrt{I(\theta)}$
- NML: ALSO RELATED TO $I(\theta)$.