http://www.cs.washington.edu/education/courses/cse446/13sp/

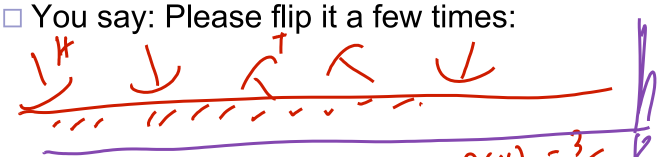
Point Estimation

Machine Learning – CSE446
Carlos Guestrin
University of Washington

April 3, 2013

Your first consulting job

- A billionaire from the suburbs of Seattle asks you a question:
  - He says: I have thumbtack, if I flip it, what's the probability it will fall with the nail up?
  - You say: Please flip it a few times:



  - You say: The probability is: $P(H) = 3/5$
  - **He says: Why???**
  - You say: Because…

Model

Thumbtack – Binomial Distribution

Observed
HHTTH

- P(Heads) = θ, P(Tails) = 1-θ

$P(D|\theta) = P(HHTTH) = \theta\,\theta\,(1-\theta)\,(1-\theta)\,\theta$
$= \theta^3\,(1-\theta)^2$

learning task
choose $\hat{\theta}$
pick θ that maximizes the likelihood of seeing these observations according to my model

- Flips are i.i.d.:
  - Independent events
  - Identically distributed according to Binomial distribution

iid

- Sequence *D* of α_H Heads and α_T Tails

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

MLE

argmax$_\theta$ f(θ) = argmax$_\theta$ ln f(θ)   ln. is monotonic

Maximum Likelihood Estimation

→ argmax is argument of max(·) function

- **Data:** Observed set *D* of α_H Heads and α_T Tails
- **Hypothesis:** Binomial distribution
- Learning θ is an optimization problem
  - What's the objective function?

$P(D|\theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$

- MLE: Choose θ that maximizes the probability of observed data:

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \; P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \; \ln P(\mathcal{D} \mid \theta)\end{aligned}$$

$\ln a \cdot b = \ln a + \ln b$ $\ln a^b = b \ln a$

Your first learning algorithm

$\frac{d}{d\theta} \ln \theta = \frac{1}{\theta}$

$\frac{d}{d\theta} \ln(1-\theta) = \frac{-1}{1-\theta}$

$$\begin{aligned} \widehat{\theta} &= \arg\max_{\theta} \ \ln P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \ \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T} \end{aligned}$$

- Set derivative to zero: $\boxed{\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) = 0}$

$\frac{d}{d\theta} \ln P(D|\theta) = \frac{d}{d\theta} [\ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}] = \frac{d}{d\theta}[\alpha_H \ln\theta + \alpha_T \ln(1-\theta)]$

$= \alpha_H \frac{d}{d\theta} \ln\theta + \alpha_T \frac{d}{d\theta} \ln(1-\theta) = \frac{\alpha_H}{\theta} - \frac{\alpha_T}{1-\theta} = 0$

$\theta = \frac{\alpha_H}{\alpha_H+\alpha_T} = \frac{3}{2+3} = \frac{3}{5}$ ✓ good!! ☺

How many flips do I need?

$$\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

- Billionaire says: I flipped 3 heads and 2 tails.
- You say: θ = 3/5, I can prove it! ← MLE
- He says: What if I flipped 30 heads and 20 tails?
- You say: Same answer, I can prove it!
- **He says: What's better?**
- You say: Humm… The more the merrier???
- He says: Is this why I am paying you the big bucks???

Simple bound (based on Hoeffding's inequality)

- For $N = \alpha_H + \alpha_T$, and $\widehat{\theta}_{MLE} = \dfrac{\alpha_H}{\alpha_H + \alpha_T}$

datapoints

- Let θ^* be the true parameter, for any $\varepsilon>0$:

accuracy within 0.1

prob $\hat{\theta}_{MLE}$ is bad estimate of θ^* ⇒ lose your job

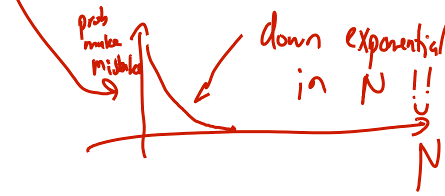
$$P(\mid \widehat{\theta}_{MLE} - \theta^* \mid \geq \epsilon) \leq 2e^{-2N\epsilon^2}$$

truth

tolerance

#times I flip coin

say we pick $\varepsilon = 0.1$



PAC Learning → Sample complexity bounds

$1-\delta$ $\quad$ ε

- PAC: Probably Approximate Correct
- Billionaire says: I want to know the thumbtack parameter θ, within $\varepsilon = 0.1$, with probability at least $1-\delta = 0.95$. How many flips?

$$P(\mid \widehat{\theta} - \theta^* \mid \geq \epsilon) \leq 2e^{-2N\epsilon^2} \leq \delta$$

δ = my tolerance to losing my job

$$\ln \delta \geq \ln 2 - 2N\varepsilon^2$$

$$N \geq \frac{\ln \frac{2}{\delta}}{2\varepsilon^2}$$

if $\delta = 0.05$
$\varepsilon = 0.1$
$\Downarrow$
$N \geq 184.4$ flips

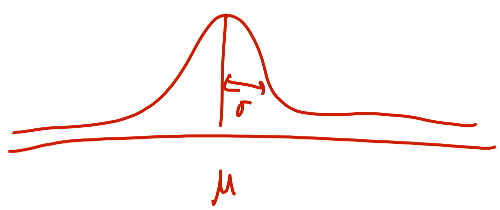
pretty bad bounds tend to be loose

What about continuous variables?

- Billionaire says: If I am measuring a continuous variable, what can you do for me? Salary of employees
- **You say: Let me tell you about Gaussians…** Normal distributions

mean Std dev

$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$



Some properties of Gaussians

- affine transformation (multiplying by scalar and adding a constant)
  - $X \sim N(\mu,\sigma^2)$
  - $Y = aX + b$ ➔ $Y \sim N(a\mu+b, a^2\sigma^2)$

    linear transform on mean

    constants
- Sum of Gaussians
  - $X \sim N(\mu_X,\sigma^2_X)$
  - $Y \sim N(\mu_Y,\sigma^2_Y)$
  - $Z = X+Y$ ➔ $Z \sim N(\mu_X+\mu_Y, \sigma^2_X+\sigma^2_Y)$

Learning a Gaussian

Exam scores

$x_1 = 85$

$x_2 = 92$

$\vdots$

$x_n = 97$

- Collect a bunch of data
  - Hopefully, i.i.d. samples
  - e.g., exam scores

why?? MLE

- Learn parameters
  - Mean $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
  - Variance σ^2

$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

MLE for Gaussian

- Prob. of i.i.d. samples $D=\{x_1,\ldots,x_N\}$:

$(x_1, \ldots, x_N)$

$$\underbrace{P(\mathcal{D} \mid \mu, \sigma)} = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^N e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

$$\mu_{MLE}, \sigma_{MLE} = \arg\max_{\mu,\sigma} P(D \mid \mu, \sigma) = \arg\max_{\mu,\sigma} \ln P(D \mid \mu, \sigma)$$

- Log-likelihood of data:

$\arg\max_{\mu,\sigma}$

$$\begin{aligned}
\ln P(\mathcal{D} \mid \mu, \sigma) &= \ln\left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^N e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}\right] \\
&= -N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^N \frac{(x_i-\mu)^2}{2\sigma^2}
\end{aligned}$$

ln prob of observing data

Your second learning algorithm: MLE for mean of a Gaussian

$\frac{d}{d\mu}(f(\mu)+g(\mu)) = \frac{df}{d\mu} + \frac{dg}{d\mu}$

- What's MLE for mean?

$$\frac{d}{d\mu} \ln P(\mathcal{D} \mid \mu, \sigma) = \frac{d}{d\mu}\left[\underbrace{\cancel{-N \ln \sigma\sqrt{2\pi}}}_{}^{0 \text{ does not depend on } \mu} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}\right] = 0$$

$$\frac{\partial \ln P(D|\mu,\sigma)}{\partial \mu} = -\sum_{i=1}^{N} \frac{\partial}{\partial \mu}\frac{(x_i-\mu)^2}{2\sigma^2} = \sum_{i=1}^{N} \frac{x_i-\mu}{\cancel{\sigma^2}} = 0$$

multiply both sides by σ^2

$$\frac{\partial}{\partial \mu}\frac{(x_i-\mu)^2}{2\sigma^2} = -\frac{x_i-\mu}{\sigma^2}$$

$$N\mu = \sum_{i=1}^{N} \mu = \sum_{i=1}^{N} x_i$$

$$\Rightarrow \hat{\mu}_{MLE} = \frac{\sum_{i=1}^{N} x_i}{N}$$

$\hat{\mu}_{MLE}$ does not depend on choice of σ

MLE for variance

- Again, set derivative to zero:

$$\frac{d}{d\sigma} \ln P(\mathcal{D} \mid \mu, \sigma) = \frac{d}{d\sigma}\left[-N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= \underbrace{\frac{d}{d\sigma}\left[-N \ln \sigma\sqrt{2\pi}\right]}_{-N/\sigma} - \sum_{i=1}^{N} \underbrace{\frac{d}{d\sigma}\left[\frac{(x_i-\mu)^2}{2\sigma^2}\right]}_{-\frac{(x_i-\mu)^2}{\sigma^3}} = 0$$

$$\Rightarrow -\frac{N}{\sigma} + \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{\sigma^3} = 0$$

use $\mu = \hat{\mu}_{MLE}$ because optimum choice of μ doesn't depend on σ

$$\Rightarrow \sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \hat{\mu}_{MLE})^2}{N}$$

Learning Gaussian parameters

you now know two learning algs binomial + Gaussians

- MLE:

$$\hat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N} (x_i - \hat{\mu})^2$$

- BTW. MLE for the variance of a Gaussian is **biased**
  - Expected result of estimation is **not** true parameter!
  - Unbiased variance estimator:

$$\hat{\sigma}^2_{unbiased} = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \hat{\mu})^2$$

What you need to know…

- Learning is…
  - Collect some data
    - E.g., thumbtack flips
  - Choose a hypothesis class or model
    - E.g., binomial
  - Choose a loss function
    - E.g., data likelihood
  - Choose an optimization procedure
    - E.g., set derivative to zero to obtain MLE
  - Collect the big bucks

this is ML

- Like everything in life, there is a lot more to learn…
  - Many more facets… Many more nuances…
  - The fun will continue…

Announcement: R Tutorial

- R: open-source scripting language for stats & ML
- A lot of resources online

- Tutorial:
  - When: Thursday April 4th at 6:00pm
  - Where: EEB 125

- Before attending please download and install R:
  - http://www.r-project.org/
- We recommend using an R environment such as:
  - R studio: http://www.rstudio.com/
  - Tinn-R: http://www.sciviews.org/Tinn-R/

subscribe to Google Group please!! :)