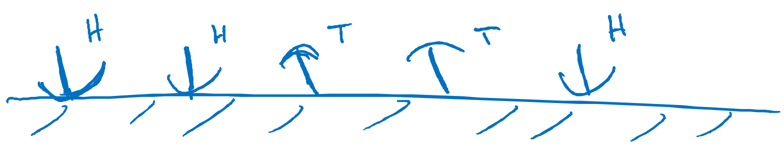
Point Estimation

CSE 446: Machine Learning
Emily Fox
University of Washington
January 6, 2017

Maximum likelihood estimation for a binomial distribution

Your first consulting job

- A bored Seattle billionaire 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:



3 H
2 T

  - You say: The probability is: $\frac{3}{5}$
  - **He says: Why???**
  - You say: Because...

Thumbtack – Binomial distribution

- P(Heads) = θ, P(Tails) = 1-θ   param
- Flips are i.i.d.:
  - Independent events
  - Identically distributed according to a binomial distribution
- Sequence D of α_H heads (H) and α_T tails (T)   p(T)
- P(D | θ) = $p(HHTTH \mid \theta) = \theta \cdot \theta \cdot (1-\theta) \cdot (1-\theta) \cdot \theta = \theta^3 (1-\theta)^2$

  cond. or given

  p(H)

  generically.. $= \theta^{\alpha_H} (1-\theta)^{\alpha_T}$ ← If I knew θ I could compute the prob. of the seq.

The learning task

binomial dist.

- Want to learn a model of thumbtack flips from experience ← data
- **Example 1:** Maximum likelihood estimation
  What value of θ maximizes the likelihood of having seen the observed sequence (according to my model)?
- What is a likelihood function?

$$\mathcal{L}(\theta) = p(\underbrace{D}_{\text{data}} \mid \underbrace{\theta}_{\text{params}})$$

function of θ

Is this a prob.? Does it add to 1 over θ? NO

Maximum likelihood estimation

- **Data:** Observed set D of α_H heads (H) and α_T tails (T) ← "experience"
- **Hypothesis:** Binomial distribution ← assumption of how the world works
- Learning θ is an optimization problem
  - What's the objective function?

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

- **MLE:** Choose θ that maximizes the likelihood of observed data

$$\hat{\theta} = \arg\max_\theta P(D \mid \theta)$$
$$= \arg\max_\theta \ln P(D \mid \theta)$$

argument of max(·) fcn

$\arg\max_\theta f(\theta) = \arg\max_\theta \ln f(x)$

ln is monotonic

Your first learning algorithm

$$\hat{\theta} = \arg\max_{\theta} \ln P(D \mid \theta)$$

$$= \arg\max_{\theta} \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T} = \arg\max_{\theta} \ln \theta^{\alpha_H} + \ln(1-\theta)^{\alpha_T} = \arg\max_{\theta} \alpha_H \ln\theta + \alpha_T \ln(1-\theta)$$

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

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

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

$$= \frac{\alpha_H}{\theta} - \frac{\alpha_T}{(1-\theta)} = 0 \rightarrow \hat{\theta} = \frac{\alpha_H}{\alpha_H + \alpha_T} = \frac{3}{3+2} = \boxed{\frac{3}{5}}$$

How many flips do I need?

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

- Billionaire says: I flipped 3 heads and 2 tails.
- You say: $\hat{\theta}$ = 3/5, I can prove it!
- He says: What if I flipped 30 heads and 20 tails?
- You say: Same answer, I can prove it! $\hat{\theta} = \frac{3}{5}$
- **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 $\hat{\theta}_{MLE} = \dfrac{\alpha_H}{\alpha_H + \alpha_T}$

total # datapoints

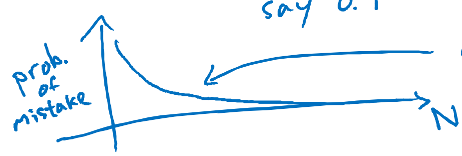
- Let θ^* be the true parameter. For any $\varepsilon>0$:

$$P(|\underbrace{\hat{\theta}_{MLE} - \theta^*}_{\text{error}}| \geq \epsilon) \leq 2e^{-2N\epsilon^2}$$

prob. $\hat{\theta}_{MLE}$ is a bad esti. of θ^* (lose your job!)

small tolerance, say 0.1

flips



goes down exp fast with N!

9 ©2017 Emily Fox CSE 446: Machine Learning

PAC learning

- **PAC:** Probably Approximate Correct
- Billionaire says: I want to know the thumbtack parameter θ within ε = 0.1, with probability at least 1-δ = 0.95. How many flips do I need?

$$P(|\hat{\theta}_{MLE} - \theta^*| \geq \epsilon) \leq \boxed{2e^{-2N\epsilon^2} \leq \delta}$$

my tolerance to losing my job

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

$$N \geq \frac{\ln 2/\delta}{2\epsilon^2}$$

If $\delta = 0.05$ $\epsilon = 0.1$

$\Downarrow$

$N \geq 184.4$ flips

pretty loose bounds in practice

10 ©2017 Emily Fox CSE 446: Machine Learning

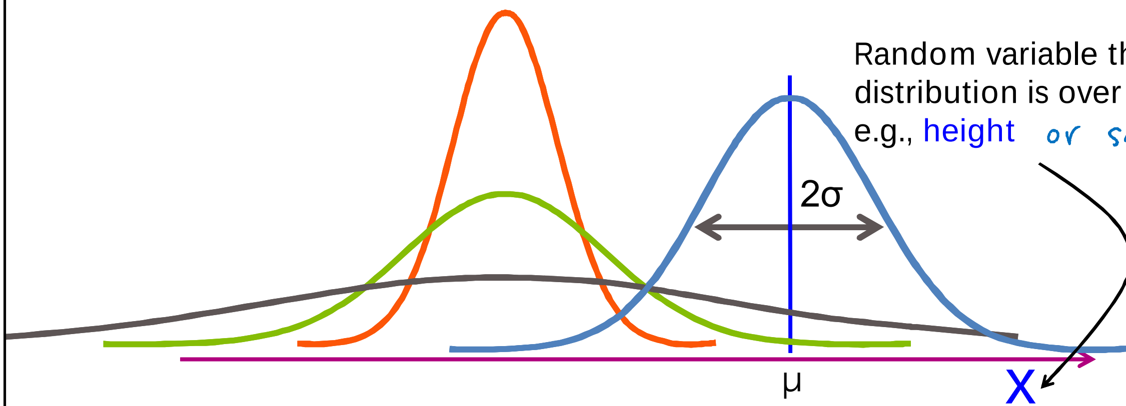
What about continuous-valued data?

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

1D Gaussians

Fully specified by **mean** μ a **variance** σ^2
(or **standard deviation** σ)



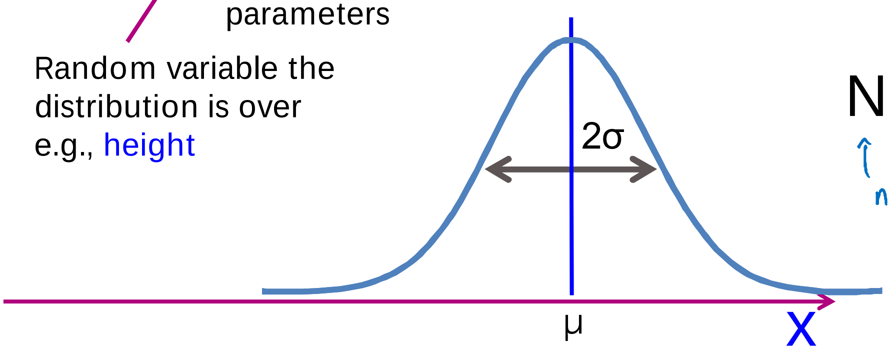
Random variable the distribution is over
e.g., height or salaries

13 ©2016 Emily Fox & Carlos Guestrin CSE 446: Machine Learning

1D Gaussian probability density function

$$p(x \mid \underbrace{\overbrace{\mu, \sigma}^{\theta}}_{\text{parameters}}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Random variable the distribution is over
e.g., height



$N(\mu, \sigma^2)$

normal dist.

14 ©2016 Emily Fox & Carlos Guestrin CSE 446: Machine Learning

Some properties of Gaussians

- Affine transformation (multiplying by scalar and adding a constant) "distributed as"
  - X ~ N(μ,σ^2)
  - Y = aX + b → Y ~ N(aμ+b,$a^2\sigma^2$)

known (deterministic) scalars

same affine trans.

$E[aX+b] = aE[X] + b = a\mu + b$

- Sum of Gaussians
  - X ~ N(μ_X,σ^2_X)
  - Y ~ N(μ_Y,σ^2_Y)
  - Z = X+Y → Z ~ N($\mu_X+\mu_Y$, $\sigma^2_X+\sigma^2_Y$)

Learning a Gaussian

- Collect a bunch of data
  - Hopefully, i.i.d. samples
  - e.g., heights of students in class
- Learn parameters
  - Mean $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$ ← why? MLE
  - Variance $\hat{\sigma}^2$

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

MLE for Gaussian

- Prob. of i.i.d. samples D={x_1,…,x_N}:

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

datapt i

$$\hat{\mu}_{MLE}, \hat{\sigma}_{MLE} = \underset{\mu,\sigma}{\arg\max}\; p(D \mid \mu, \sigma) = \underset{\mu,\sigma}{\arg\max}\; \ln p(D \mid \mu, \sigma)$$

- Log-likelihood of data:

$$\ln p(D \mid \mu, \sigma) = \ln\left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right]$$

$$= -N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}$$

Your second learning algorithm: MLE for mean of a Gaussian

- What's MLE for the mean?

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

no depend. on μ

$$\frac{d}{d\mu} \ln p(D \mid \mu, \sigma) = -\sum_{i=1}^{N} \frac{d}{d\mu} \frac{(x_i-\mu)^2}{2\sigma^2} = \sum_{i=1}^{N} \frac{x_i - \mu}{\cancel{\sigma^2}} = 0$$

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

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

MLE doesn't depend on choice of σ^2

MLE for variance

- Again, set derivative to zero:

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

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

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

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

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

Learning Gaussian parameters

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

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

- FYI, MLE for the variance of a Gaussian is **biased**
  - Expected value of estimator is **not** true parameter!
  - Unbiased variance estimator:

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

Recap of concepts

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

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