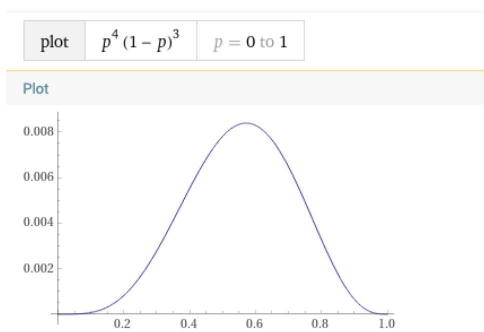
CAS CS 365
Lec 11

Probability & Machine Learning

1. Overview
    a. Maximum Likelihood Estimation → parameter model (pdf, pmf)
        i. Max L(D| θ) → example θ = (μ, o^2) when it is normal
        ii. L(D; θ) = multiply from i =1 to i=n (p^x_i) * (1-p)^(1-x_i)
            1. X_i = 1 if the ith toss is H, 0 otherwise
            2. P: Heads: $\sum x_i$
            3. 1-P: Tails: $\sum 1\text{-}x_i$
    b. In other words, we assume iid samples from some distribution
2. The Thumbtack problem (cont.)
    a. We shall denote Pr(HTHHTHT) as Pr(HTHHTHT;p)
        i. P is not a random variable
    b. D data, θ parameter
    c. We refer to Pr(D|θ) as the likelihood of the data under the model
    d. In our problem, Pr(D;p) = p^4 (1-p)^3
3. Maximum Likelihood Estimate (MLE)
    a. Data: thumbtack tosses
    b. Hypothesis: A flip is a Bernoulli distributed variable. Independence of flips
    c. Learning: Find p* that maximizes the data likelihood

$$\Pr(\mathcal{D};p) = p^4(1-p)^3$$

    d. 

4. Optimize to find p*
    a. f(p) = p^4(1-p)^3
    b. Max f(p)
    c. Take the logarithm of L(D;p) to make life easier → log(L(D;p)) since the product becomes the sum
    d.

$$p^* = \arg\max_{p} \log\left(p^{a_H}(1-p)^{a_T}\right)$$

$$\frac{d}{dp}(*) = \frac{d}{dp}\left(a_H \log p + a_T \log(1-p)\right) =$$

$$= a_H \frac{1}{p} + a_T\left(-\frac{1}{1-p}\right) = 0 \Rightarrow$$

$$\Rightarrow a_H/p = a_T/1-p \Rightarrow p a_T = a_H(1-p) \Rightarrow$$

$$\Rightarrow p(\overbrace{a_H + a_T}^{n}) = a_H \Rightarrow \boxed{p_{MLE} = \frac{a_H}{n}}$$

    e. P = 4/7
5. Confidence intervals (reminder)
    a. Boston billionaire says: I want to know the true p within 0.01 accuracy, with confidence at least 95%
    b. Sampling theorem: Given n independent 0-1 RVs X_i such that $\Pr(X_i = 1) = p(i=1,\ldots,n)$ where $n \geq \frac{3}{\epsilon^2}\ln\left(\frac{2}{\delta}\right)$ then the following holds:

$$\Pr\left(\left|\frac{\sum_{i=1}^{n} X_i}{n} - p\right| \leq \epsilon\right) \geq 1 - \delta$$

6. Two important properties of the MLE
    a. Consistent
        i. $\theta_{\text{MLE}} \rightarrow \theta_{\text{true}}$ in probability
    b. Equivariant
        i. If θ_{MLE} is the MLE of $\theta_{\text{true}} \Rightarrow g(\theta_{\text{MLE}})$ is the MLE of $g(\theta_{\text{true}})$

ii. This means that if we want to the function it (example: p^2), we simply put the function
iii. For example, p_{hat}^2 for p^2

7. Billionaire with prior beliefs
    a. He says: Wait! I know that the thumbtack should be close to 50-50
    b. You say: let's be Bayesian
    c. Rather than learn a single value for p, we learn a probability distribution
        i. P now becomes a random variable
8. Inference using Bayes' rule
    a. Notice the notation Pr(D|p) instead of Pr(D;p)
    b. $$Pr(p|D) = \frac{Pr(D|p)Pr(p)}{Pr(D)}$$
    c. Pr(D) → does not depend on p
9. Bayesian inference - summary
    a. We choose the prior distribution f(θ). This distribution expresses our prior beliefs on the parameter θ.
    b. We choose the statistical model for the likelihood function f(D|θ)
    c. After observing the data D=X_1,...,X_n, we update our beliefs and calculate the posterior distribution f(θ|D)
    d. Maximum a posteriori (MAP) estimate is the mode of the posterior distbrution
10. Important observation
    a. If we impose a uniform prior on p, then
    $$Pr(p|D) \propto Pr(D|p)$$
    b. Image denoising lecture
        i. Had we imposed a uniform prior on images x, then our MAP inference would be the same as the MLE
        ii. Choosing a good prior is important in applications of Bayesian inference
11. Conjugate priors
    a. Definition
        i. "If the posterior distribution p(θ|x) is in the same probability distribution family as the prior probability distribution p(θ), the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function p(x|θ)" → assume that prior is same as likelihood
    b. Conjugate prior: create fictitious data
        i. For example: billionaire is sure that the probability of head to tails is 8:2
        ii. Append 800^H, 200^T (the more confident you are, the more data you collect) from the data H^4, T^3, you append the data
        iii. Therefore, H → 804, T → 203, so P(H) = 804/1007

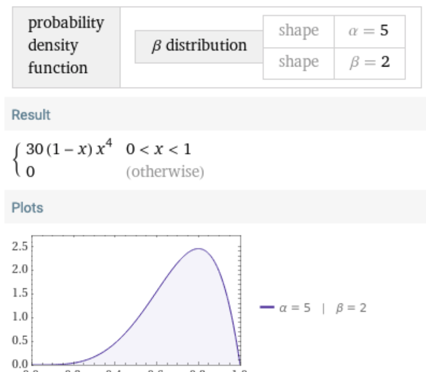
12. Bayesian inference for the thumbtack problem

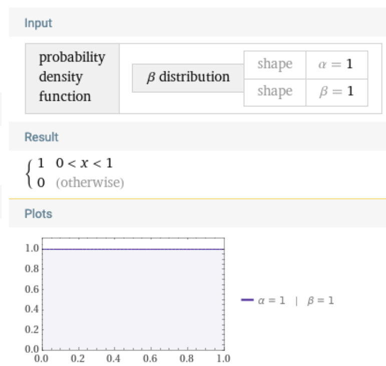
a. The probability density for beta distribution is

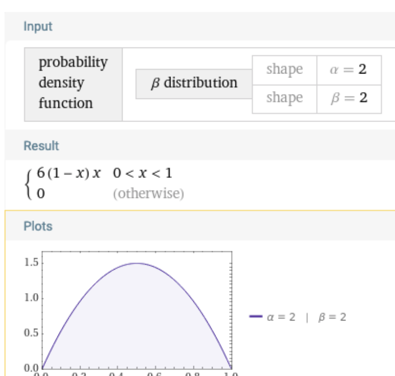
$$f(x;a,b)=\frac{\Gamma(a+b)x^{a-1}(1-x)^{b-1}}{\Gamma(a)\Gamma(b)},\ 0\le x\le 1, a,b>0$$

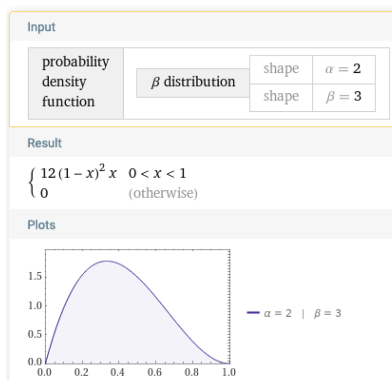
where $\Gamma(z)=\int_0^\infty x^{z-1}e^{-x}\,dx$ is the gamma function

b.









c. As you increase beta and theta, you increase the confidence

$$P(p) \propto p^{\beta_H-1}(1-p)^{\beta_T-1}$$ PRIOR

$$P(D|p) = p^{a_H}(1-p)^{a_T}$$ LIKELIHOOD.

POSTERIOR

$$P(p|D) \propto P(D|p)\,P(p) \propto p^{a_H+\beta_H-1}(1-p)^{a_T+\beta_T-1}$$

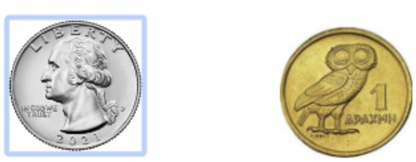
$$= \text{Beta}(a_H+\beta_H,\ a_T+\beta_T)$$

fictitious coin tosses reflecting our prior belief.

d.

13. Method of moments
    a. Suppose our model has parameters $\boldsymbol{\theta}=(\theta_1,..,\theta_k)$
        i. Recall that the j-th moment of a RV X is E[X^j]. To denote that this is a function of the model, we write $E_{\boldsymbol{\theta}}[X^j]$.
            1. Compute analytically, $\boxed{\alpha_j = \mathbb{E}_\theta\left[X^j\right],\ j = 1,\ldots,k}$
        ii. Consider the j-th sample moment for data data x_1, …, x_n is

$$\boxed{\hat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} x_i^j,\ j = 1,\ldots,k} = \mu_{MOM}$$

        iii. $\sum x_i$^2/n = o^2$_{MOM}$
        iv. Equate the analytical moment expressions with the sample moments, and solve a system of k equations with unknowns to learn $\boldsymbol{\theta}=(\theta_1,..,\theta_k)$
14. Method of methods: example 1
    a. Let X1, …, Xn ~ Bernoulli(p)
        i. We have one parameter, so k=1.
        ii. The first moment is the mean $\alpha_1=E_p[X]=p$.
        iii. The first sample moment is the sample mean.
        iv. Thus we directly get $p_{\text{MoM}}= \frac{1}{n}\sum_{i=1}^{n} x_i$
        v. Remark: Here, MOM is same as MLE, but this is not always the case
15. MLE vs MAP
    a. MLE
        i. Goal: Find θ maximizing the log-likelihood Pr(x;θ) → also denoted as Pr(x|θ)
    b. MAP
        i. Goal: Find θ maximizing the posterior Pr(θ|x)
    c. In some cases, solving analytically for θ is hard
    d. EM algorithm is an iterative approach to solving hard parameter learning problems
16. Two coin problem
    a. 

$\Pr(H) = \theta_A$ $\quad\quad$ $\Pr(H) = \theta_B$

    b. Process
        i. Suppose we choose a coin (A or B) uniformly at random
        ii. We toss the coin n times, and record the total number of heads

c. We repeat the process k times
d. We have two unknown parameters: θ_A, θ_B
e. Suppose k = 5, n = 10
f. The data are two vectors
    i. $x=X_H = (x_1, x_2, x_3, x_4, x_5)$ #heads per round
    ii. $z=Z_c = (z_1, z_2, z_3, z_4, z_5)$ #coin id per round
g. E.g., x=(5,9,8,4,7), z=(B, A, A, B, A)
h. Solve two separate maximum posterior
i.

Coin A	Coin B
	5 H, 5 T
9 H, 1 T	
8 H, 2 T	
	4 H, 6 T
7 H, 3 T	
24 H, 6 T	9 H, 11 T

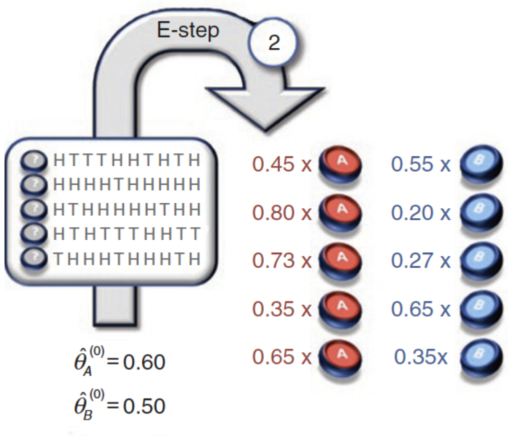
$$\hat{\theta}_A = \frac{24}{24+6} = 0.80$$

$$\hat{\theta}_B = \frac{9}{9+11} = 0.45$$

j. Use EM if there is missing data
k. Suppose we only see the number of heads per round
    i. In other words, we do not have access to z
l. We refer to z as hidden variables or latent factors
m. Remarks
    i. Not uncommon/common setting in data science applications → missing data
    ii. Clear that maximizing $\Pr(x|\theta)$ is much harder than $\Pr(x,z|\theta)$ in the presence of missing data z
n. We will proceed iteratively by updating
o. Each iteration starts with a guess of the unknown parameters

$$\hat{\theta}^{(t)} = \left(\hat{\theta}_A^{(t)}, \hat{\theta}_B^{(t)}\right)$$

p. E-step: a probability distribution over possible completions is computed using the current parameters $\hat{\theta}^{(t)}$
q. M-step: the new parameters are determined using the current completions
r. Suppose our initial guess for the unknown variables are 0.6, 0.5
s. E-step: what is the probability that round i comes from coin A/ coin B?

t. 

u. Likelihoods

Coin A: $\propto 0.6^5 \cdot 0.4^5 \approx 0.000796$

Coin B $\propto 0.5^5\, 0.5^5 = \left(\frac{1}{2}\right)^{10} \approx 0.000976$

$Pr(Z_1 = A) = \frac{0.6^5 \cdot 0.4^5}{0.6^5 0.4^5 + (0.5)^{10}} = 0.45, \quad Pr(Z_1 = B) = 0.55$

v. M-step: in order to learn $\hat{\theta}^{(t+1)} = \left(\hat{\theta}_A^{(t+1)}, \hat{\theta}_B^{(t+1)}\right)$, we first need to estimate the number of heads/tails from coins A/B given our estimate of the latent variables

i. Notice that instead of being 100% certain whether a round was due to coin A or B, we have a probability distribution

	Coin A	Coin B
Heads	0.45x5	0.55x5
Tails	0.45x5	0.55x5

Coin A	Coin B
≈ 2.2 H, 2.2 T	≈ 2.8 H, 2.8 T

We repeat this for all five rounds

w.

x. Suppose we do one more round with different number of heads and tails

One more round

Round 2: 9H, 1T
From the E-round we have

0.80 x (A)   0.20 x (B)

	Coin A	Coin B
Heads	0.8x9	0.2x9
Tails	0.8x1	0.2x1

≈ 7.2 H, 0.8 T   ≈ 1.8 H, 0.2 T

Likelihoods Round 2

Coin A: $\propto 0.6^9 \cdot 0.4^1 \approx 0.00403$

Coin B $\propto 0.5^9 \, 0.5^1 = \left(\frac{1}{2}\right)^{10} \approx 0.000976$

$Pr(Z_2 = A) = \frac{0.6^9 \cdot 0.4^1}{0.6^9 0.4^1 + (0.5)^{10}} = 0.8049 \approx 0.8, \; Pr(Z_2 = B) = 0.2$

y.

z. Having done this for all 5 rounds, we obtain the following

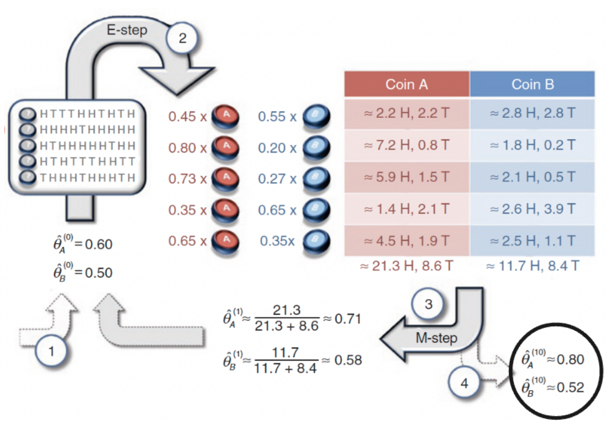
Coin A	Coin B
≈ 2.2 H, 2.2 T	≈ 2.8 H, 2.8 T
≈ 7.2 H, 0.8 T	≈ 1.8 H, 0.2 T
≈ 5.9 H, 1.5 T	≈ 2.1 H, 0.5 T
≈ 1.4 H, 2.1 T	≈ 2.6 H, 3.9 T
≈ 4.5 H, 1.9 T	≈ 2.5 H, 1.1 T
≈ 21.3 H, 8.6 T	≈ 11.7 H, 8.4 T

aa. The M-step now simply becomes the MLEs according to this data

$$\hat{\theta}_A^{(1)} \approx \frac{21.3}{21.3 + 8.6} \approx 0.71$$

$$\hat{\theta}_B^{(1)} \approx \frac{11.7}{11.7 + 8.4} \approx 0.58$$

bb. Expectation maximization



17. EM algorithm

a. Suppose maximizing $\Pr(x|\theta)$ has no closed form solution/intractable
b. Key idea: latent variables z that make likelihood computations tractable
c. Intuition: $\Pr(x,z|\theta)$, $\Pr(z|x,\theta)$ should be easy to compute after introducing the "right" latent variables
d. EM guaranteed to converge to a local maximum
e. Define "expected log" $Q(\theta|\theta')$ where θ' is the current estimate of θ:

$$Q(\theta \mid \theta') = \sum_z \Pr(z \mid x, \theta') \log \Pr(x, z \mid \theta)$$

f. The EM algorithm is an iterative method consisting of two steps
    i. E-step: Find $Q(\theta|\theta')$ in terms of the latent variables z
    ii. M-step: find θ^* maximizing $Q(\theta|\theta')$