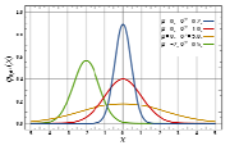
CSE 446: Point Estimation
Winter 2012

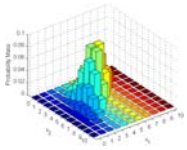
Dan Weld

Slides adapted from Carlos Guestrin (& Luke Zettlemoyer)

Random Variable

- Discrete or Continuous
  - **Boolean:** like an atom in probabilistic logic
    - Denotes some event (eg, die will come up "6")
  - **Continuous:** domain is R numbers
    - Denotes some quantity (age of a person taken from CSE446)
- (Joint) Probability Distribution





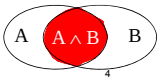
Bayesian Methods

- Learning and classification methods based on probability theory.
  - Uses ***prior*** probability of each category
    Given no information about an item.
  - Produces a ***posterior*** probability distribution over possible categories
    Given a description of an item.

- Bayes theorem plays a critical role in probabilistic learning and classification.

Axioms of Probability Theory

- All probabilities between 0 and 1
  $0 \le P(A) \le 1$
- Probability of truth and falsity
  P(true) = 1 P(false) = 0.
- The probability of disjunction is:
  $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

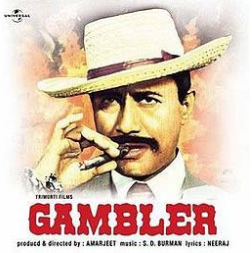


De Finetti (1931)

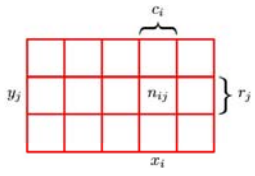
An agent who bets according to probabilities that violate these axioms…

Can be forced to bet so as to ***lose money***

Regardless of the outcome



Terminology



Marginal Probability

$p(X = x_i) = \frac{c_i}{N}$

Joint Probability

$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$

Conditional Probability

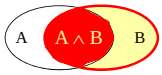
$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$

X value is given

Conditional Probability

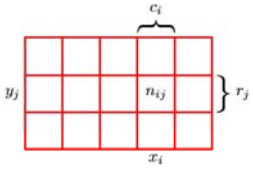
- P(*A* | *B*) is the probability of *A* given *B*
- Assumes:
  - *B* is all and only information known.
- Defined by:

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$



7

Probability Theory



Sum Rule

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N}\sum_{j=1}^{L} n_{ij}$$

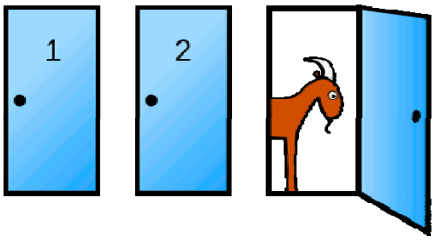
$$= \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

Product Rule

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N}$$

$$= p(Y = y_j \mid X = x_i)p(X = x_i)$$

Monty Hall

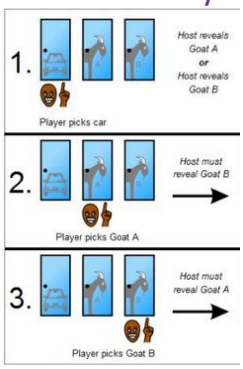


9

Monty Hall

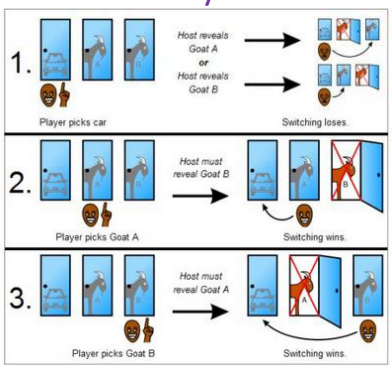
10

Monty Hall



11

Monty Hall



12

Independence

- *A* and *B* are *independent* iff:

$$P(A \mid B) = P(A)$$
$$P(B \mid A) = P(B)$$

These constraints are logically equivalent

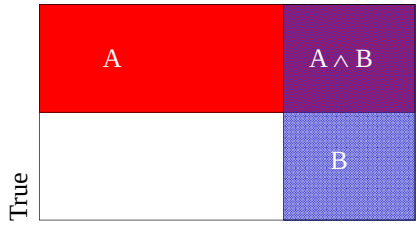
- Therefore, if *A* and *B* are independent:

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)} = P(A)$$

$$P(A \wedge B) = P(A)P(B)$$

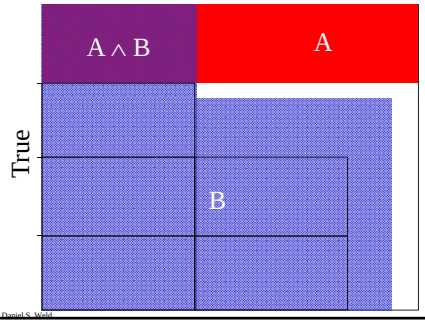
Independence

$P(A \wedge B) = P(A)P(B)$



Independence Is Rare

A&B *not* independent, since P(A|B) ≠ P(A)



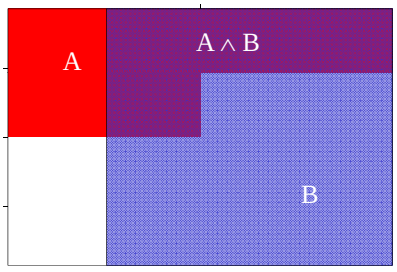
P(A) = 25%

P(B) = 80%

P(A|B) ≤ 1/7

Conditional Independence….?

Are A & B independent? P(A|B) ≤ P(A)



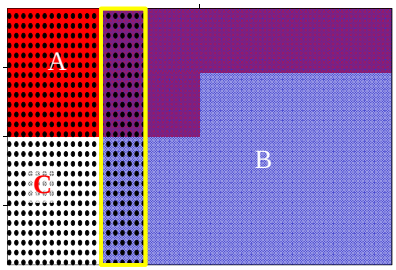
P(A)=(.25+.5)/2 = .375

P(B)= .75

P(A|B)=(.25+.25+.5)/3 =.3333

A, B Conditionally Independent Given C

P(A|B,C) = P(A|C) C = spots



P(A|C) =.5
P(B|C) = .333
P(A|B,C)=.5

Bayes Theorem

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}$$

Simple proof from definition of conditional probability:

$$P(H \mid E) = \frac{P(H \wedge E)}{P(E)} \quad \text{(Def. cond. prob.)}$$

$$\frac{P(H \wedge E)}{P(H)} = P(E \mid H) \quad \text{(Def. cond. prob.)}$$

$$P(H \wedge E) = P(E \mid H)P(H) \quad \text{(Mult both sides of 2 by P(H).)}$$

QED:

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)} \quad \text{(Substitute 3 in 1.)}$$

Medical Diagnosis

Encephalitis Computadoris

Extremely Nasty

Extremely Rare 1/100M

DNA Testing 0% false negatives

1/10,000 false positive

Scared? $P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}$

P(E)? In 100M people: 1+ 10,000

P(H|E)=1*0.00000001 / 0.00010001 ~ 0.000099

Your first consulting job

Billionaire (eccentric) Eastside tech founder asks:

- 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…

Thumbtack – Binomial Distribution

- P(Heads) = θ, P(Tails) = 1-θ

- Flips are *i.i.d.*:
  - Independent events
  - Identically distributed according to Binomial distribution
- Sequence *D* of α_H Heads and α_T Tails

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

Maximum Likelihood Estimation

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

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

- **MLE:** Choose θ to maximize probability of *D*

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

Your first parameter learning algorithm

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

- Set derivative to zero, and solve!

$$\begin{aligned}\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) &= \frac{d}{d\theta}\left[\ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}\right] \\ &= \frac{d}{d\theta}\left[\alpha_H \ln\theta + \alpha_T \ln(1-\theta)\right] \\ &= \alpha_H \frac{d}{d\theta}\ln\theta + \alpha_T \frac{d}{d\theta}\ln(1-\theta) \\ &= \frac{\alpha_H}{\theta} - \frac{\alpha_T}{1-\theta} = 0\end{aligned}$$

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

But, how many flips do I need?

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

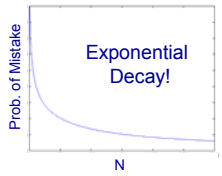
- Billionaire says: I flipped 3 heads and 2 tails.
- You say: θ = 3/5, I can prove it!
- He says: What if I flipped 30 heads and 20 tails?
- You say: Same answer, I can prove it!
- **He says: What's better?**
- You say: Umm… The more the merrier???
- He says: Is this why I am paying you the big bucks???

A bound (from Hoeffding's inequality)

For $N = \alpha_H + \alpha_T$, and $\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$

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

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



PAC Learning

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

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

$$\delta \geq 2e^{-2N\epsilon^2} \geq P(\text{mistake})$$

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

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

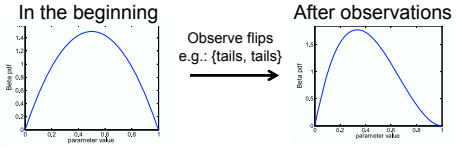
Interesting! Lets look at some numbers!

ε = 0.1, δ=0.05

$$N \geq \frac{\ln(2/0.05)}{2 \times 0.1^2} \approx \frac{3.8}{0.02} = 190$$

What if I have prior beliefs?

- Billionaire says: Wait, I know that the thumbtack is "close" to 50-50. What can you do for me now?
- **You say: I can learn it the Bayesian way…**
- Rather than estimating a single θ, we obtain a distribution over possible values of θ



Bayesian Learning

Use Bayes rule:

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta) P(\theta)}{P(\mathcal{D})}$$

Data Likelihood

Prior

Posterior

Normalization

Or equivalently: $P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta) P(\theta)$

Also, for uniform priors:

→ reduces to MLE objective

$$P(\theta) \propto 1 \qquad P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)$$

Bayesian Learning for Thumbtacks

$$P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta) P(\theta)$$

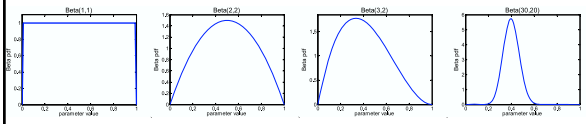
Likelihood function is Binomial:

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

- What about prior?
  - Represent expert knowledge
  - Simple posterior form
- Conjugate priors:
  - Closed-form representation of posterior
  - **For Binomial, conjugate prior is Beta distribution**

Beta prior distribution – P(θ)

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$



- Likelihood function: $P(\mathcal{D} \mid \theta) = \theta^{\alpha_H} (1-\theta)^{\alpha_T}$
- Posterior: $P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta) P(\theta)$

$$P(\theta \mid \mathcal{D}) \propto \theta^{\alpha_H} (1-\theta)^{\alpha_T} \; \theta^{\beta_H - 1} (1-\theta)^{\beta_T - 1}$$

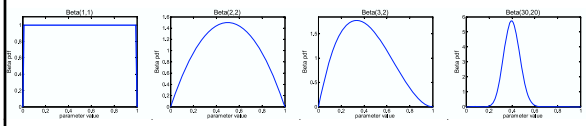
$$= \theta^{\alpha_H + \beta_H - 1} (1-\theta)^{\alpha_T + \beta_t + 1}$$

$$= Beta(\alpha_H + \beta_H, \alpha_T + \beta_T)$$

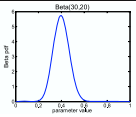
Posterior Distribution

- Prior: $Beta(\beta_H, \beta_T)$
- Data: α_H heads and α_T tails
- Posterior distribution:

$$P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$



Bayesian Posterior Inference



- Posterior distribution:

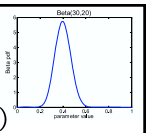
$$P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

- Bayesian inference:
  - No longer single parameter
  - For any specific f, the function of interest
  - Compute the expected value of f

$$E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid \mathcal{D}) d\theta$$

  - Integral is often hard to compute

MAP: Maximum a Posteriori Approximation



$$P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

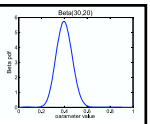
$$E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid \mathcal{D}) d\theta$$

- As more data is observed, Beta is more certain
- MAP: use most likely parameter to approximate the expectation

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

$$E[f(\theta)] \approx f(\hat{\theta})$$

MAP for Beta distribution



$$P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H + \alpha_H - 1}(1-\theta)^{\beta_T + \alpha_T - 1}}{B(\beta_H + \alpha_H, \beta_T + \alpha_T)} \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

MAP: use most likely parameter:

$$\hat{\theta} = \arg\max_\theta P(\theta \mid \mathcal{D}) = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

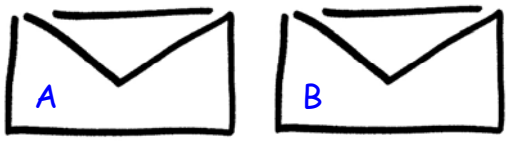
Beta prior equivalent to extra thumbtack flips

As $N \to \infty$, prior is "forgotten"

But, for small sample size, prior is important!

Envelope Problem

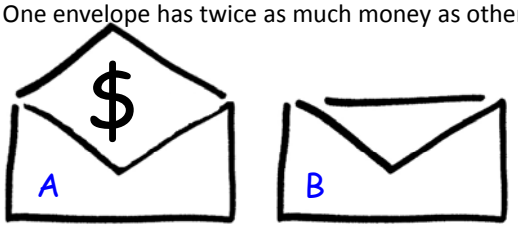
One envelope has twice as much money as other



Switch?

Envelope Problem

One envelope has twice as much money as other



A has $20

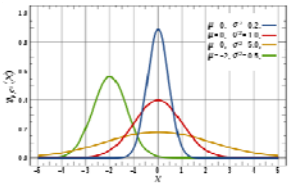
E[B] = $ ½ (10 + 40)

= $ 25

Switch?

What about continuous variables?

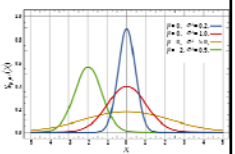
- Billionaire says: If I am measuring a continuous variable, what can you do for me?
- **You say: Let me tell you about Gaussians…**



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

Some properties of Gaussians

- Affine transformation (multiplying by scalar and adding a constant) are Gaussian
  - $X \sim N(\mu, \sigma^2)$
  - $Y = aX + b \rightarrow Y \sim N(a\mu+b, a^2\sigma^2)$
- Sum of Gaussians is Gaussian
  - $X \sim N(\mu_X, \sigma^2_X)$
  - $Y \sim N(\mu_Y, \sigma^2_Y)$
  - $Z = X+Y \rightarrow Z \sim N(\mu_X+\mu_Y, \sigma^2_X+\sigma^2_Y)$
- Easy to differentiate, as we will see soon!



Learning a Gaussian

- Collect a bunch of data
  - Hopefully, i.i.d. samples
  - e.g., exam scores
- Learn parameters
  - Mean: μ
  - Variance: σ

X_i $i =$	Exam Score
0	85
1	95
2	100
3	12
…	…
99	89

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

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

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

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

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

- Log-likelihood of data:

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

Your second learning algorithm: MLE for mean of a Gaussian

- What's MLE for mean?

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

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

MLE for variance

- Again, set derivative to zero:

$$\frac{d}{d\sigma} \ln P(\mathcal{D} \mid \mu, \sigma) = \frac{d}{d\sigma}\left[-N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}\right]$$
$$= \frac{d}{d\sigma}\left[-N \ln \sigma\sqrt{2\pi}\right] - \sum_{i=1}^{N} \frac{d}{d\sigma}\left[\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$
$$= -\frac{N}{\sigma} + \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{\sigma^3} = 0$$

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

Learning Gaussian parameters

- MLE:

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

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

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

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

Bayesian learning of Gaussian parameters

- Conjugate priors
  - Mean: Gaussian prior
  - Variance: Wishart Distribution
- Prior for mean:

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

MAP for mean of Gaussian

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

$$\frac{d}{d\mu}[\ln P(\mathcal{D} \mid \mu) P(\mu)] = \frac{d}{d\mu}[\ln P(\mathcal{D} \mid \mu) + \ln P(\mu)]$$