CAP 5610: Machine Learning

Lecture 2:
Review of Probability Theory

Instructor: Dr. Gita Sukthankar
Email: gitars@eecs.ucf.edu

Reference Materials From: GuoJun Qi, Xi Chen, Ray Mooney

Agenda

- Quick history of machine learning
- Probability theory
- Bayes
- MLE and MAP
- Example applications

History of Machine Learning Methods

1950s

–Samuel's checker player

–Selfridge's Pandemonium (vision)

1960s:

–Neural networks: Perceptron

–Pattern recognition

–Learning in the limit theory

–Minsky and Papert prove limitations of perceptron for XOR

1970s:

–Symbolic concept induction

–Winston's arch learner

–Expert systems and the knowledge acquisition bottleneck

–Quinlan's ID3 (decision tree)

–Scientific discovery with BACON

–Mathematical discovery (Lenat)

History of Machine Learning Methods

1980s:

–Advanced decision tree and rule learning

–Explanation-based Learning (EBL)

–Utility problem

–Analogy

–Cognitive architectures

–Resurgence of neural networks (connectionism, backpropagation)

–Valiant's PAC Learning Theory

–Focus on experimental methodology

History of Machine Learning Methods

1990s:

–Data mining

–Text learning

–Probabilistic models

–Reinforcement learning (RL)

–Inductive Logic Programming (ILP)

–Ensembles: Bagging, Boosting, and Stacking

2000s:

–Support vector machines

–Kernel methods

–Graphical models

–Statistical relational learning

–Transfer learning

–Sequence labeling

–Collective classification and structured outputs

2010: Deep Learning

Recap : Machine learning problem

- A set of training example $\{(x_i, y_i) | i=1,\ldots,n\}$ where x_i is the feature vector of attributes for an example i, and y_i is its label
- A set of hypotheses H
- A machine learning algorithm aims to solve an optimization problem (sample mean of error on training set)

$$h^* = \min_{h \in H} \frac{1}{n} \sum_{i=1}^{n} err(h(x_i), y_i)$$

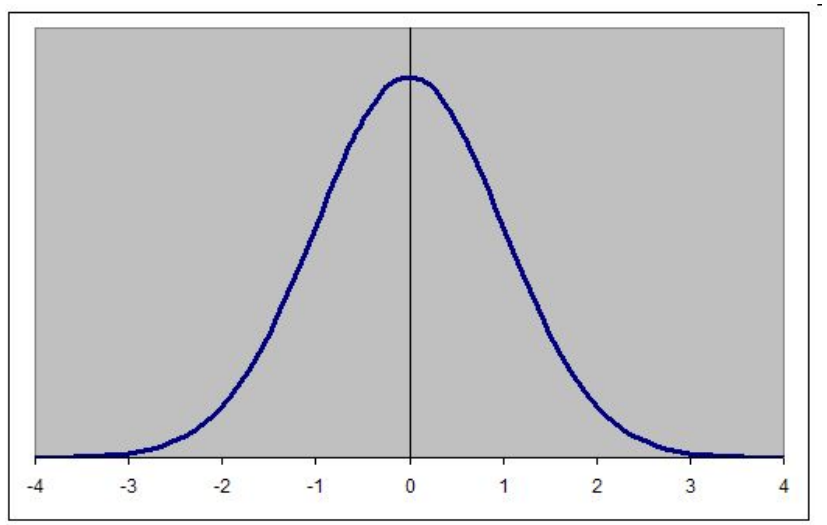
- The above problem is an approximation to the population mean

$$h^* = \min_{h \in H} E_{x \sim D} err\big(h(x), h_o(x)\big)$$

Probability

- Population mean

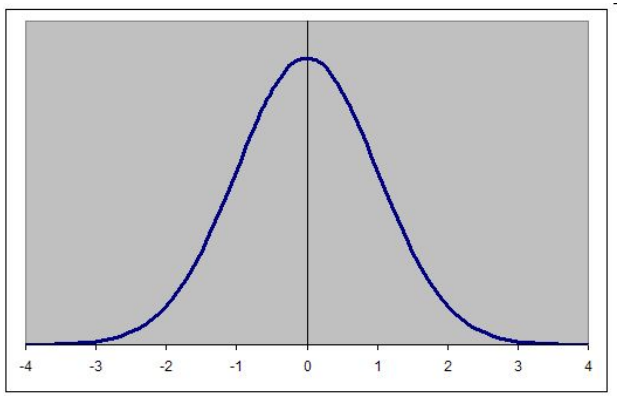
$$h^* = \min_{h \in H} E_{x \sim D} err\big(h(x), h_o(x)\big)$$



Distribution D

- Data x is a random variable drawn from a distribution D
- A distribution is a mathematical model which depicts the likelihood that a random variable x is drawn from.

Distribution – Continuous case



- Probability density p(x)
- The larger the p(x) for a variable x, the more likely that a variable around x will be generated by this distribution.
- The probability that x falls into an interval [a,b] can be computed as

$$\int_a^b p(x)dx$$

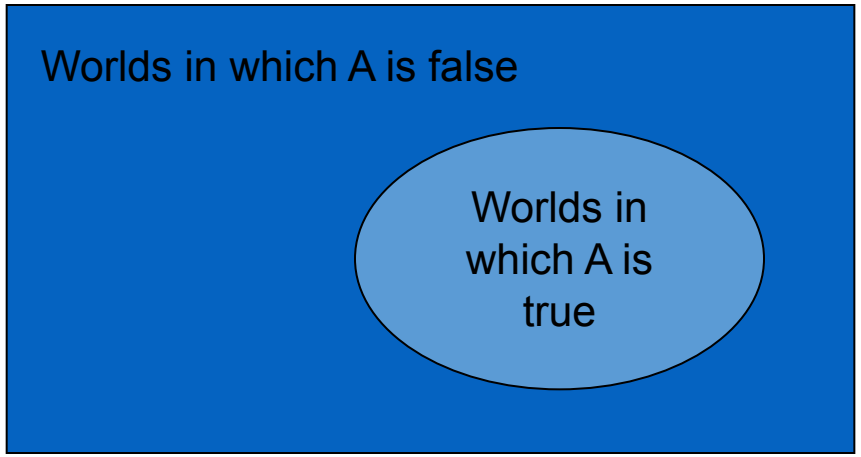
Properties of continuous distribution

- Nonnegative: $p(x) \geq 0$
- The probability that any variable drawn from p(x) must fall between positive and negative infinity

$$\int_{-\infty}^{+\infty} p(x)dx=1$$

Distribution for discrete case

- A discrete random variable is defined on a discrete set
  - E.g., A is a random variable drawn from Ω={1,2,3,4,5,6}, denoting the outcome of tossing a dice
- A probability p(A) is a function that maps a discrete value A onto the interval [0,1].
- P(A) is also called the probability measure or mass of A.



P(A) is the area of the oval

Distribution for discrete case

- The probability that A falls into a subset S of Ω is

$$P(\Omega) = \sum_{x \in \Omega} P(A = x) = 1$$

$$P(S) = \int_{x \epsilon S} p(A = x) dx = \sum_{x \epsilon S} p(A = x)$$

Discrete Distributions

- Bernoulli distribution: Ber(*p*) (special case: binomial n=1)

$$P(x) = \begin{cases} 1-p & \text{for } x = 0 \\ p & \text{for } x = 1 \end{cases} \qquad P(x) = p^x(1-p)^{1-x}$$

- Binomial distribution: Bin(n,p)
  - Suppose a coin with head prob. *p* is tossed *n* times.
  - What is the probability of getting *k* heads?
  - How many ways can you get *k* heads in a sequence of *k* heads and *n-k* tails?

$$\Pr(K = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Statistical Characterizations

- **Expectation:** the centre of mass, mean, first moment

$$E(X) = \sum_{i \in S} = x_i p(x_i) \quad \text{(discrete)}$$

$$E(X) = \int_{-\infty}^{\infty} x p(x) dx \quad \text{(continuous)}$$

- **Sample mean:**

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

- **Variance:** the spread

$$Var(X) = \sum_{x \in S} [x_i - E(X)]^2 p(x_i) \quad \text{(discrete)}$$

$$Var(X) = \int_{\infty}^{\infty} [x - E(X)]^2 p(x) dx \quad \text{(continuous)}$$

- **Sample variance:**

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

Note that the sample variance is divided by N-1 to avoid underestimates.

Implications for Machine Learning

- Population mean

$$h^* = \min_{h \in H} E_{(x,y)\sim D} err(h(x), y) = \int err(h(x), y) p(x, y) dx dy$$

- Is the average error over all possible variables (input data) drawn from a certain joint distribution p(x,y) of input x and output label y.
- P(x,y) is unknown in general, so we do some experiments which generate a set of training examples to approximate it.

With training set $\{(x_i, y_i) | i=1,\ldots,n\}$, we have sample mean:

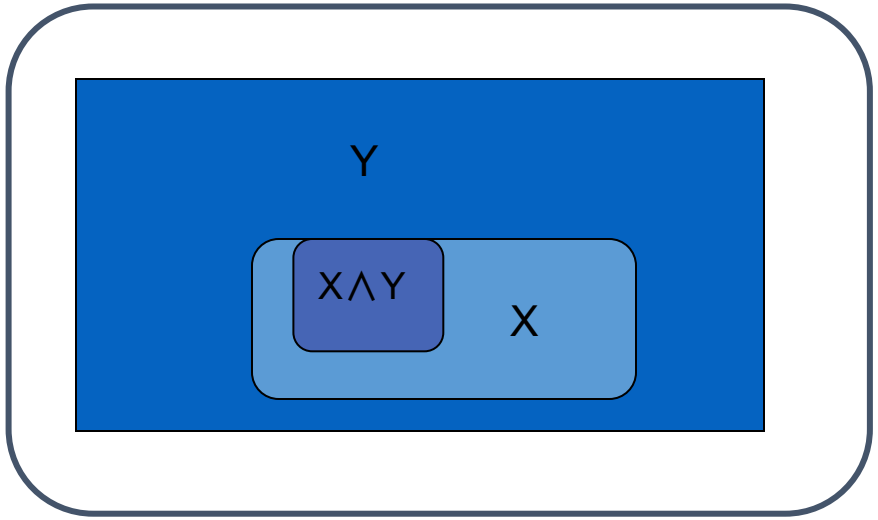
$$h^* = \min_{h \in H} \frac{1}{n} \sum_{i=1}^{n} err(h(x_i), y_i)$$

Conditional Probability

- P(X|Y) = the likelihood of getting X if Y is generated
  - H = "having a headache"
  - F = “infected with Flu"
    - P(H)=1/10
    - P(F)=1/40
    - P(H|F)=1/2
  - P(H|F) = fraction of headache given you have a flu
    = P(H∧F)/P(F)
- Definition:

$$P(X|Y) = \frac{P(X \wedge Y)}{P(Y)}$$

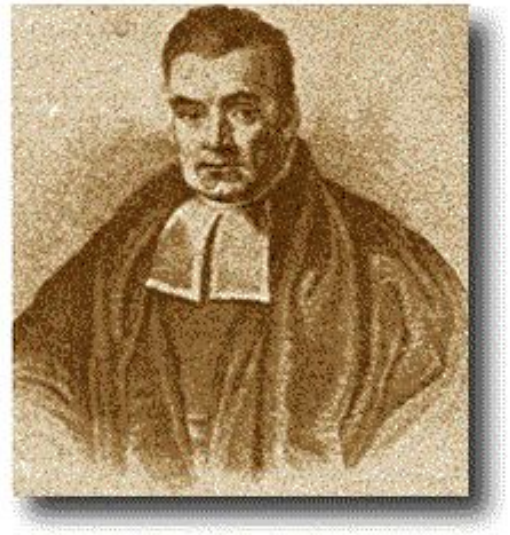
$$P(X \wedge Y) = P(X|Y)P(Y)$$



Bayes Rule

- What we have just did leads to the following general expression:

$$P(Y|X) = \frac{P(X|Y)p(Y)}{P(X)}$$



Rev. Thomas Bayes

This is Bayes Rule

$$P(Y = y_i|X) = \frac{P(X|Y = y_i)p(Y = y_i)}{\sum_{j \in S} P(X|Y = y_j)P(Y = y_j)}$$

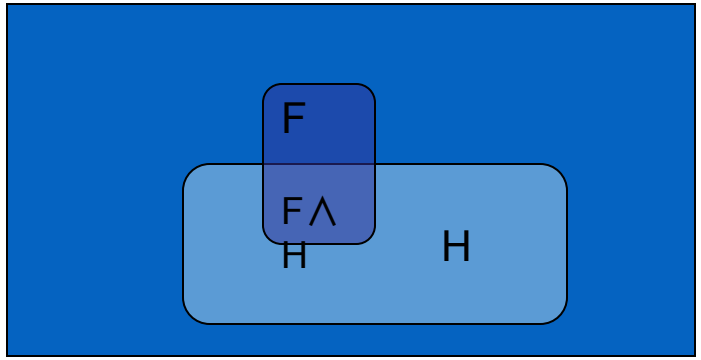
Probabilistic Inference

- H = "having a headache"
- F = “infected with flu"
  - P(H)=1/10
  - P(F)=1/40
  - P(H|F)=1/2

Note: İt is more useful to know whether you have the flu given the observation of a headache but easier to measure from data whether a headache was observed in a set of known flu patients

- The Problem:

P(F|H) = ?



Joint Probability

- A joint probability distribution for a set of random variables gives the probability of every value
  - **P**(*Flu,DrinkBeer*) = a 2 × 2 matrix of values:

	B	¬B
F	0.005	0.02
¬F	0.195	0.78

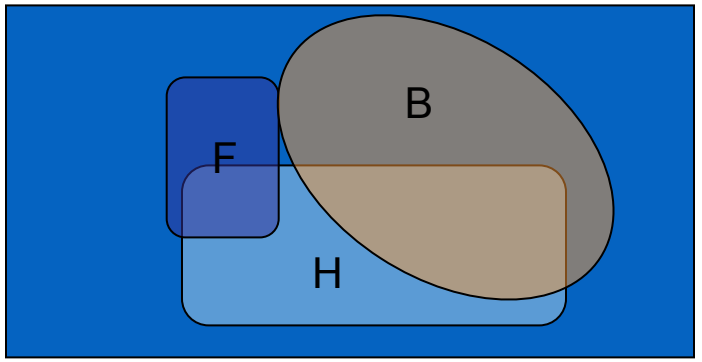
Every question about a domain can be answered by the joint distribution but it requires an exponential amount of parameters to specify.

Inference with the Joint

- Compute Marginals

$$P(\text{flu} \cap \text{headache})0.05 + 0.015 = 0.065$$

¬F	¬B	¬H	0.4
¬F	¬B	H	0.1
¬F	B	¬H	0.17
¬F	B	H	0.2
F	¬B	¬H	0.05
F	¬B	H	0.05
F	B	¬H	0.015
F	B	H	0.015

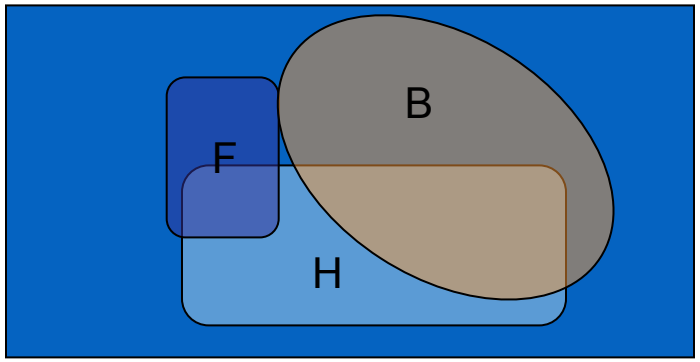


Inference with the Joint

- Compute Marginals

$$P(\text{Headache}) = 0.1 + 0.2 + 0.05 + 0.015 = 0.365$$

¬F	¬B	¬H	0.4
¬F	¬B	H	0.1
¬F	B	¬H	0.17
¬F	B	H	0.2
F	¬B	¬H	0.05
F	¬B	H	0.05
F	B	¬H	0.015
F	B	H	0.015



Inference with the Joint

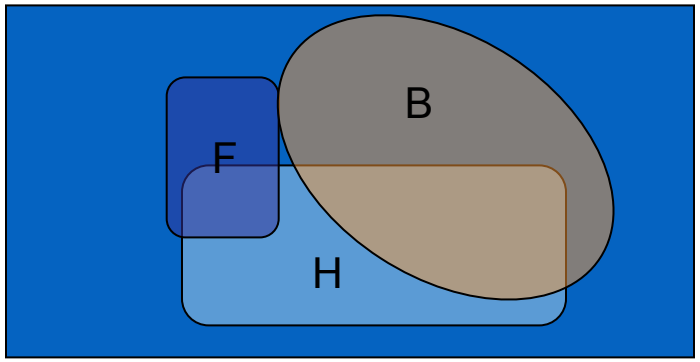
- Compute Conditionals

$$P(\text{flu}|\text{headache}) = \frac{P(\text{flu} \cap \text{headache})}{P(\text{headache})}$$

$$= \frac{0.065}{0.365} = 0.178$$

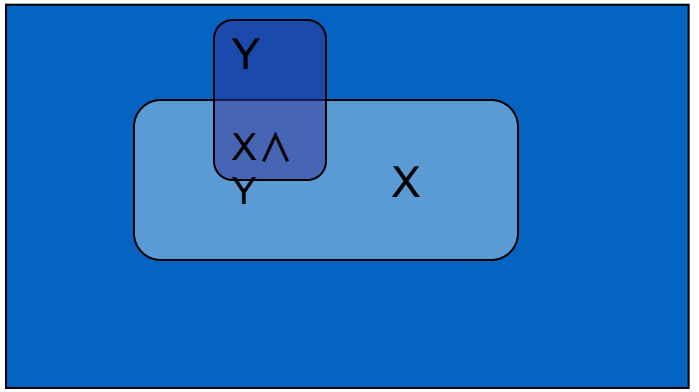
¬F	¬B	¬H	0.4
¬F	¬B	H	0.1
¬F	B	¬H	0.17
¬F	B	H	0.2
F	¬B	¬H	0.05
F	¬B	H	0.05
F	B	¬H	0.015
F	B	H	0.015

General idea: compute distribution on output variable by **fixing** input **evidence variables** and **summing** over **hidden variables**



Conditional Independence

- Random variables X and Y are said to be independent if:
  - $P(X \cap Y) = P(X)*P(Y)$
- Alternatively, this can be written as
  - $P(X \mid Y) = P(X)$ and
  - $P(Y \mid X) = P(Y)$
- Intuitively, this means that telling you that the fact that Y happened, does not make X more or less likely.
- Note: This does not mean X and Y are disjoint!!!



Marginal and Conditional Independence

- Recall that for events E (i.e. $X=x$) and H (say, $Y=y$), the conditional probability of E given H, written as $P(E|H)$, is:

  $P(E \text{ and } H)/P(H)$

  = the probability of both E and H are true, given H is true)

- E and H are (statistically) independent if

  $$P(E) = P(E|H)$$

  (i.e., prob. E is true doesn't depend on whether H is true); or equivalently

  $$P(E \text{ and } H)=P(E)P(H).$$

- E and F are *conditionally* independent given H if

  $$P(E,F|H) = P(E|H)P(F|H)$$

Parameter Learning from iid data

- Goal: for a probability density p(x, $\boldsymbol{\theta}$) with parameter $\boldsymbol{\theta}$, estimating its parameters $\boldsymbol{\theta}$ from a dataset of N independent, identically distributed (*iid*), examples

$$D = \{x_1, \ldots, x_N\}$$

- Maximum likelihood estimation (MLE)
  1. One of the most common estimators
  2. With iid and full-observability assumption, write $L(\theta)$ as the likelihood of the data

  3. pick the setting of parameters most likely to have generated the data we saw:

$$\begin{aligned} L(\theta) &= P(x_1, x_2, \ldots, x_N; \theta) \\ &= P(x_1; \theta) P(x_2; \theta), \ldots, P(x_N; \theta) \\ &= \prod_{i=1}^{N} P(x_i; \theta) \\ \theta^* &= \arg\max_{\theta} L(\theta) = \arg\max_{\theta} \log L(\theta) \end{aligned}$$

Example 1: Bernoulli model

- Data:
  - We observed N ***iid*** coin tossing: D={1, 0, 1, …, 0}
- Representation:

  Binary random variable: $x_n = \{0, 1\}$
- Model: $P(x) = \begin{cases} 1-p & \text{for } x = 0 \\ p & \text{for } x = 1 \end{cases} \Rightarrow \quad P(x) = \theta^x (1-\theta)^{1-x}$
- How to write the likelihood of a single example x_i ? $P(x_i) = \theta^{x_i}(1-\theta)^{1-x_i}$
- The likelihood of dataset D={x_1, …,x_N}:

$$P(x_1, x_2, \ldots, x_N|\theta) = \prod_{i=1}^{N} P(x_i|\theta) = \prod_{i=1}^{N} (\theta^{x_i}(1-\theta)^{1-x_i})$$

$$= \theta^{\sum_{i=1}^{N} x_i} (1-\theta)^{\sum_{i=1}^{N} 1-x_i} = \theta^{\#\text{heads}} (1-\theta)^{\#\text{ tails}}$$

MLE

- Objective function – log likelihood:

$$l(\theta; D) = \log P(D|\theta) = \log \theta^{n_h}(1-\theta)^{n_t} = n_h \log \theta + (N - n_h)\log(1-\theta)$$

- We need to maximize this w.r.t. θ

- Take derivatives wrt θ

$$\frac{\partial l}{\partial \theta} = \frac{n_h}{\theta} - \frac{N - n_h}{1-\theta} = 0 \quad \Rightarrow \quad \hat{\theta}_{MLE} = \frac{n_h}{N} \quad \text{or} \quad \hat{\theta}_{MLE} = \frac{1}{N}\sum_i x_i$$

Frequency as sample mean

Overfitting

- Recall that for Bernoulli Distribution, we have $\hat{\theta}_{ML}^{\text{head}} = \frac{n^{\text{head}}}{n^{\text{head}} + n^{\text{tail}}}$
- What if we tossed too few times and observed zero heads?
- We will predict that the probability of seeing a head next is zero!!! $\hat{\theta}_{ML}^{\text{head}} = 0$
- The rescue:
  - Where n' is known as the pseudo- (imaginary) count

$$\hat{\theta}_{ML}^{\text{head}} = \frac{n^{\text{head}} + n'}{n^{\text{head}} + n^{\text{tail}} + n'}$$

But we can do this in a more formal way….

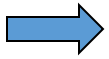
Example 2: Univariate Normal Distribution

- Data:
  - We observed N ***iid*** real samples:

    D={-0.1, 10, 1, -5.2, …, 3}
- Model: $P(x) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-(x-\mu)^2 / 2\sigma^2\right\}$
- Log likelihood: $l(\theta; D) = \log P(D \mid \theta) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{n=1}^{N}\frac{(x_n-\mu)^2}{\sigma^2}$
- MLE: take derivative and set to zero:

$$\frac{\partial l}{\partial \mu} = (1/\sigma^2)\sum_n (x_n - \mu)$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_n (x_n-\mu)^2$$



$$\mu_{MLE} = \frac{1}{N}\sum_n (x_n)$$

$$\sigma^2_{MLE} = \frac{1}{N}\sum_n (x_n - \mu_{ML})^2$$

The Bayesian Theory

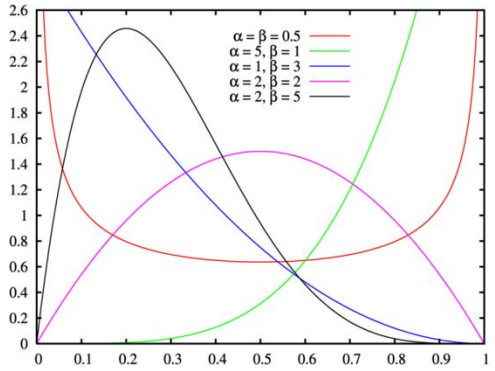
- The Bayesian Theory: (e.g., for data *D* and model parameter)
  - $p(\theta|D) = p(D|\theta)p(\theta)/p(D)$
  - θ are the parameters for the likelihood $p(x|\theta)$
  - α are the parameters for the prior $p(\theta|\alpha)$.
  - the **posterior** equals to the **likelihood** times the **prior**, up to a constant.

- This allows us to capture uncertainty about the model in a principled way
- MAP requires explicitly modeling the prior distribution of the model whereas MLE does not.

Beta distribution as a prior for Bernoulli

- Beta distribution:

$$P(\theta;\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}=B(\alpha,\beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}$$



Posterior distribution of θ :

$$P(\theta\mid x_1,\ldots,x_N)=\frac{p(x_1,\ldots,x_N\mid\theta)p(\theta)}{p(x_1,\ldots,x_N)}\propto\theta^{n_h}(1-\theta)^{n_t}\times\theta^{\alpha-1}(1-\theta)^{\beta-1}=\theta^{n_h+\alpha-1}(1-\theta)^{n_t+\beta-1}$$

- Notice the similarity of the posterior as the prior,
- Such a prior is called a **conjugate prior** to the likelihood (when the prior and posterior distributions are from related families)

Bayesian estimation for Bernoulli, cont'd

- Posterior distribution of θ :

$$P(\theta \mid x_1,\ldots,x_N) = \frac{p(x_1,\ldots,x_N \mid \theta)p(\theta)}{p(x_1,\ldots,x_N)} \propto \theta^{n_h}(1-\theta)^{n_t} \times \theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{n_h+\alpha-1}(1-\theta)^{n_t+\beta-1}$$

- Maximum *a posteriori* (MAP) estimation:

$$\theta_{MAP} = \arg\max_{\theta} \log P(\theta \mid x_1,\ldots,x_N)$$

- Posterior mean estimation:

Beta parameters can be understood as pseudo-counts

- Prior strength: A=α+β
  - A can be interpreted as the size of an imaginary data set from which we obtain the **pseudo-counts**

$$\theta_{Bayes} = \int \theta p(\theta \mid D)d\theta = C\int \theta \times \theta^{n_h+\alpha-1}(1-\theta)^{n_t+\beta-1}d\theta = \frac{n_h+\alpha}{N+\alpha+\beta}$$

Bayesian estimation for normal distribution

Gaussian is self-conjugate which makes it convenient!

- Normal Prior:

$$P(\mu) = \left(2\pi\tau^2\right)^{-1/2} \exp\left\{-(\mu-\mu_0)^2/2\tau^2\right\}$$ (model parameter)

- Joint Probability:

$$P(D,\mu) \propto P(D\,|\,\mu)P(\mu) = 2\pi\sigma^2 \quad ^{-N/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N} x_n - \mu \quad ^2\right\}$$

$$\times\ 2\pi\tau^2 \quad ^{-1/2} \exp\ -(\mu-\mu_0)^2/2\tau^2$$

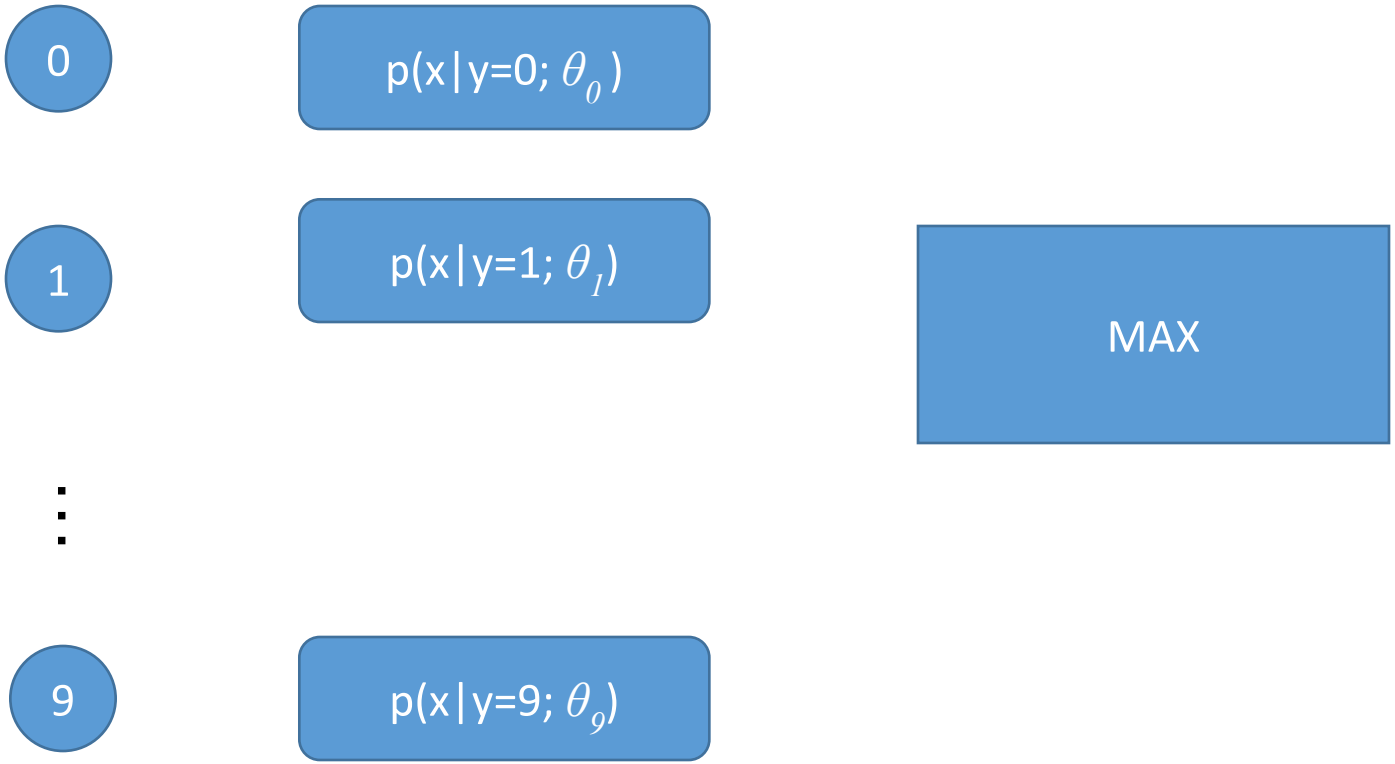
- Posterior:

$$P(\mu\,|\,x) = \left(2\pi\tilde{\sigma}^2\right)^{-1/2} \exp\left\{-(\mu-\tilde{\mu})^2/2\tilde{\sigma}^2\right\}$$

where $\tilde{\mu} = \dfrac{N/\sigma^2}{N/\sigma^2+1/\tau^2}\bar{x} + \dfrac{1/\tau^2}{N/\sigma^2+1/\tau^2}\mu_0$, and $\tilde{\sigma}^2 = \left(\dfrac{N}{\sigma^2}+\dfrac{1}{\tau^2}\right)^{-1}$

Sample mean

Parameter Estimation for Machine Learning



- Training algorithm: Estimate an individual probability distribution with the training examples of a digit.
  - Assume the distribution is Gaussian
- Prediction algorithm:
  - Given a test example x, apply each probability distribution to compute the corresponding likelihood
  - Compare the likelihoods, and predict the test example as the one with the maximal likelihood.

Practical Application

Worker quality modeling in crowdsourcing

Problem

- Use Mechanical Turk to label data
- Can use majority vote across multiple labelers to resolve dissension
- Some of the workers are good labelers and some are bad ones----good labelers have a higher probability of generating a good answer

Idea

Use worker quality model to weight the value of the labels

KPark

Crowdsourcing parking finding

"Improving the Performance of Mobile Phone Crowdsourcing Applications" Erfan Davami, Gita Sukthankar. Proceedings of the International Conference on Autonomous Agents and Multi-agent Systems. 2015. pp. 145-153.

https://www.youtube.com/watch?v=47ln3mJ9KSs

Reference Reading

- Chap 2 and Appendix B of PRML
- The information that may useful through the course
  - Continuous random variable: probability density
  - Discrete random variable: probability mass
  - Conditional probability and Bayes rule
  - Posterior and prior distribution
  - Gaussian distribution, Bernoulli distribution