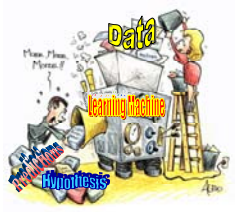
Advanced Machine Learning

Introduction



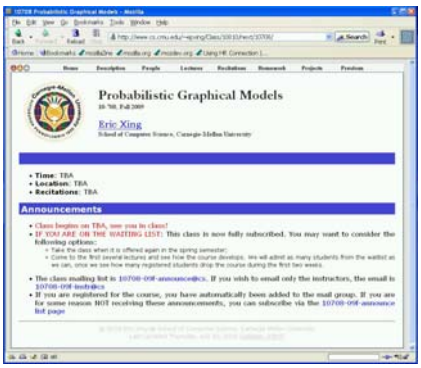
Eric Xing

Lecture 1, August 10, 2009

Reading:

Advanced Machine Learning

- Where does it come from:
  - http://www.cs.cmu.edu/~epxing/Class/10701/
  - http://www.cs.cmu.edu/~epxing/Class/10708/



Logistics

- Text book
  - Chris Bishop, **Pattern Recognition and Machine Learning (required)**
  - Tom Mitchell, **Machine Learning**
  - David Mackay, **Information Theory, Inference, and Learning Algorithms**
  - Daphnie Koller and Nir Friedman, **Probabilistic Graphical Models**
- Class resource
  - http://www.cis.pku.edu.cn/faculty/vision/wangliwei/DS.html
  - http://www.cs.cmu.edu/~epxing/Class/DS
- TA:
  - Dr. Jun Zhu
  - …
- Host:
  - Liwei Wang, Peking University
  - Changshui Zhang, Tsinghua University

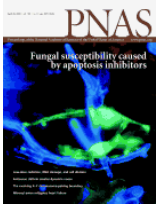
What is Learning

Learning is about seeking a predictive and/or executable understanding of natural/artificial subjects, phenomena, or activities from …

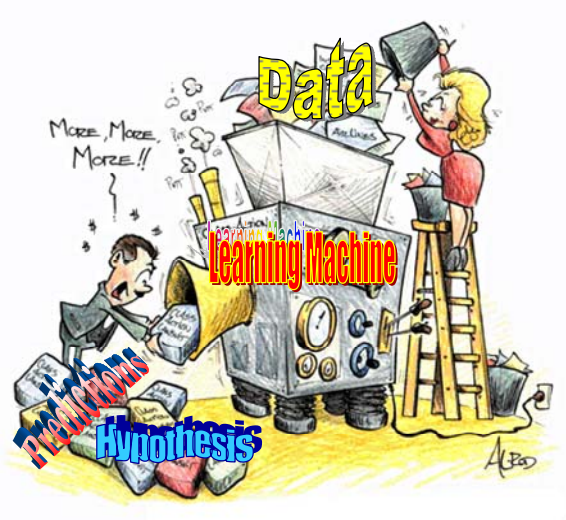
Apoptosis + Medicine

Grammatical rules
Manufacturing procedures
Natural laws
…

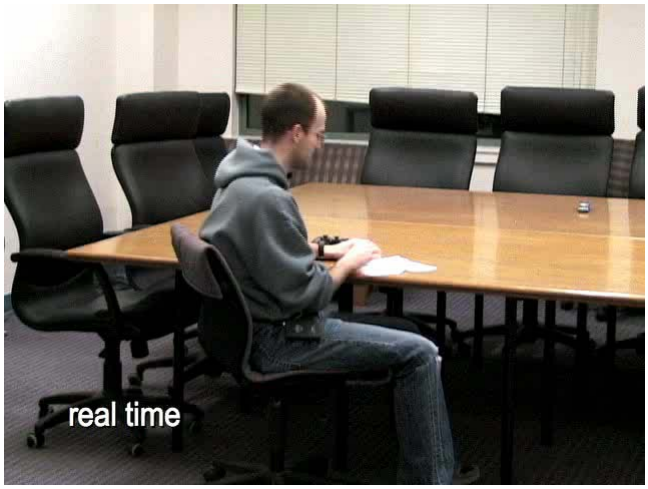
Inference



Machine Learning



Fetching a stapler from inside an office --- the Stanford STAIR robot



Machine Learning

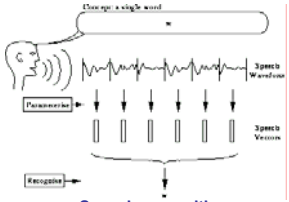
Machine Learning seeks to develop theories and computer systems for

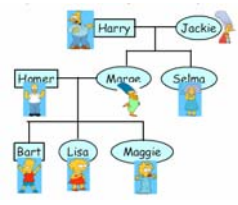
- representing;
- classifying, clustering and recognizing;
- reasoning under uncertainty;
- predicting;
- and reacting to
- …

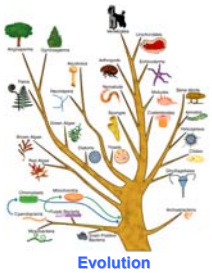
complex, real world data, based on the system's own experience with data, and (hopefully) under a unified model or mathematical framework, that

- can be formally characterized and analyzed
- can take into account human prior knowledge
- can generalize and adapt across data and domains
- can operate automatically and autonomously
- and can be interpreted and perceived by human.

Where Machine Learning is being used or can be useful?







Natural language processing and speech recognition

- Now most pocket **Speech Recognizers** or **Translators** are running on some sort of learning device --- the more you play/use them, the smarter they become!

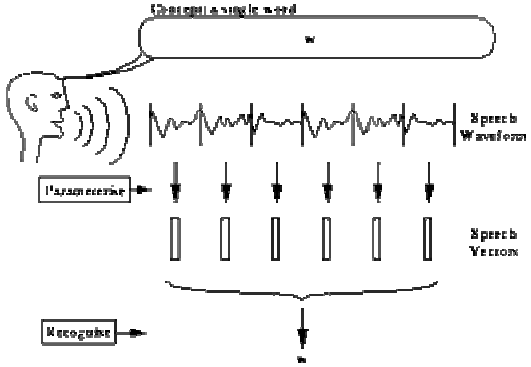


Fig. 1.2 Isolated Word Problem

Object Recognition

- Behind a security camera, most likely there is a computer that is learning and/or checking!

Robotic Control I

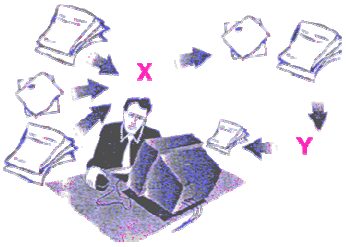
- The **best** helicopter pilot is now a computer!
  - it runs a program that learns how to fly and make acrobatic maneuvers by itself!
  - no taped instructions, joysticks, or things like …

Robotic Control II

- Now cars can find their own ways!

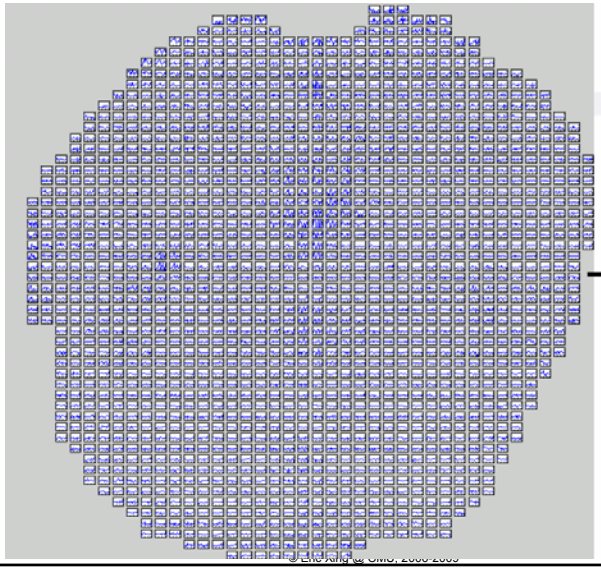
Text Mining

- Reading, digesting, and categorizing a vast text database is too much for human!



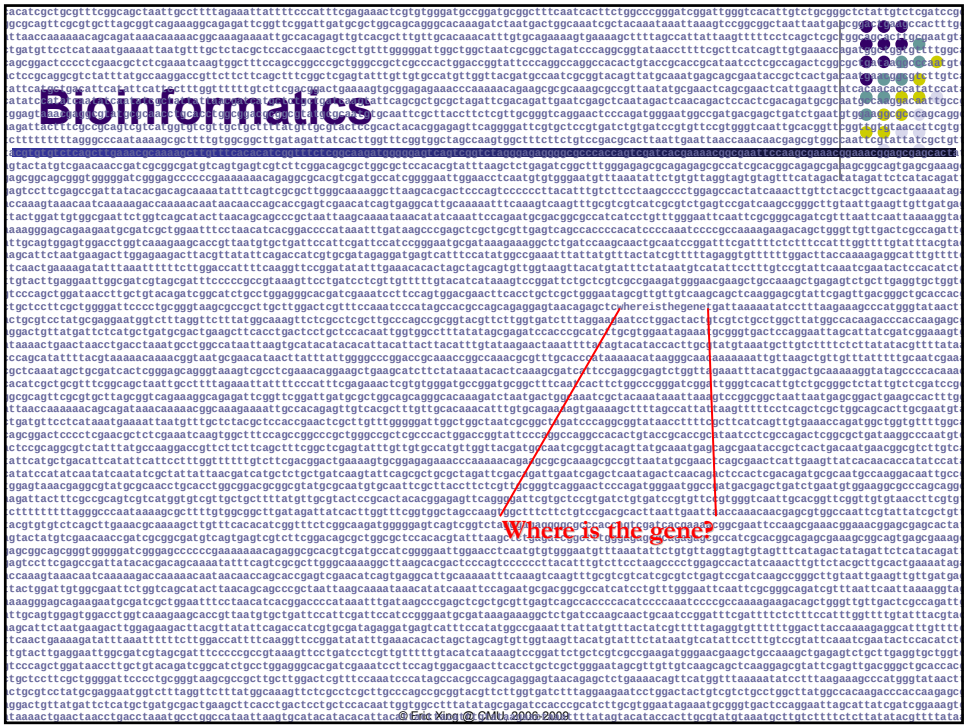
- **We want:**

Understanding Brain Activities



Reading a noun (vs verb)

[Rustandi et al., 2005]



Paradigms of Machine Learning

- Supervised Learning
  - Given $\mathcal{D} = \{\mathbf{X}_i, \mathbf{Y}_i\}$, learn $f(\cdot): \mathbf{Y}_i = f(\mathbf{X}_i)$, s.t. $\mathcal{D}^{\text{new}} = \{\mathbf{X}_j\} \Rightarrow \{\mathbf{Y}_j\}$
- Unsupervised Learning
  - Given $\mathcal{D} = \{\mathbf{X}_i\}$, learn $f(\cdot): \mathbf{Y}_i = f(\mathbf{X}_i)$, s.t. $\mathcal{D}^{\text{new}} = \{\mathbf{X}_j\} \Rightarrow \{\mathbf{Y}_j\}$
- Reinforcement Learning
  - Given $D = \{\text{env}, \text{actions}, \text{rewards}, \text{simulator/trace/real game}\}$

    learn $\begin{matrix}\text{policy}: e, r \rightarrow a \\ \text{utility}: a, e \rightarrow r\end{matrix}$, s.t. $\{\text{env}, \text{new real game}\} \Rightarrow a_1, a_2, a_3 \ldots$
- Active Learning
  - Given $\mathcal{D} \sim G(\cdot)$, learn $\mathcal{D}^{\text{new}} \sim G'(\cdot)$ and $f(\cdot)$, s.t. $\mathcal{D}^{\text{all}} \Rightarrow G'(\cdot), \text{policy}, \{\mathbf{Y}_j\}$

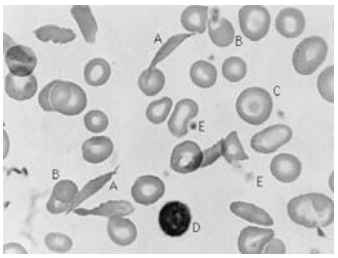
Theories of Learning

For the learned F(; θ)

- Consistency (value, pattern, …)
- Bias versus variance
- Sample complexity
- Learning rate
- Convergence
- Error bound
- Confidence
- Stability
- …

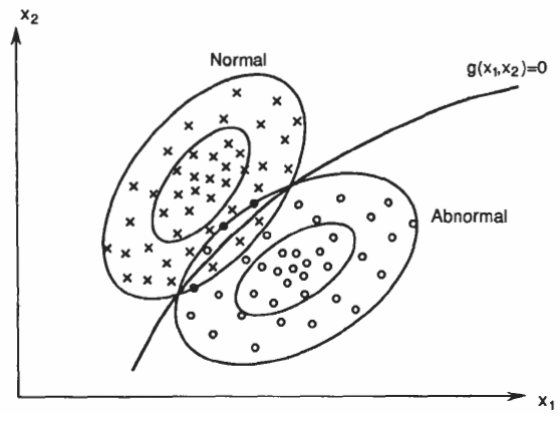
Function Approximation

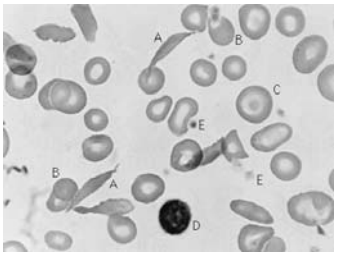
- **Setting**:
  - Set of possible instances X
  - Unknown target function $f: X \rightarrow Y$
  - Set of function hypotheses $H=\{ h \mid h: X \rightarrow Y \}$
- **Given**:
  - Training examples $\{<x_i, y_i>\}$ of unknown target function f
- **Determine**:
  - Hypothesis $h \in H$ that best approximates f



Decision-making as high-dimensional inference

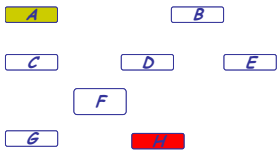
- Distributions of samples from normal and abnormal machine





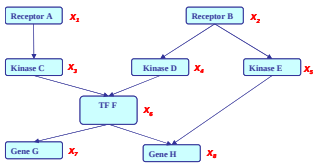
Multivariate statistical models

- What is a "model"?



$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8,)$$

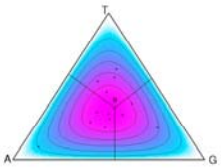
- Graphical models:



$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$
$$= P(X_1)\, P(X_2)\, P(X_3|\, X_1)\, P(X_4|\, X_2)\, P(X_5|\, X_2)\, P(X_6|\, X_3, X_4)\, P(X_7|\, X_6)\, P(X_8|\, X_5, X_6)$$

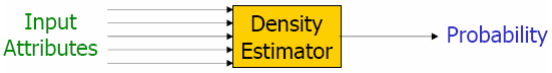
Inference
Prediction
Decision-Making under uncertainty
…

→ **Statistical Machine Learning**
→ **Function Approximation:** $F(\ |\theta)$?
→ **Density Estimation**



Density Estimation

- A Density Estimator learns a mapping from a set of attributes to a Probability



- Often know as *parameter estimation* if the distribution form is specified
  - Binomial, Gaussian …
- Three important issues:
  - Nature of the data (iid, correlated, …)
  - Objective function (MLE, MAP, …)
  - Algorithm (simple algebra, gradient methods, EM, …)
  - Evaluation scheme (likelihood on test data, predictability, consistency, …)

Discrete Distributions

- Bernoulli distribution: Ber(p)

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

- Multinomial distribution: Mult(1, θ)
  - Multinomial (indicator) variable:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{bmatrix}, \quad \text{where} \quad X_j = [0,1], \text{ and } \sum_{j\in\{1,\ldots,6\}} X_j = 1$$

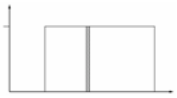
$$X_j = 1 \text{ w.p. } \theta_j, \quad \sum_{j\in\{1,\ldots,6\}} \theta_j = 1 \ .$$

$$p(x(j)) = P(\{X_j = 1, \text{where } j \text{ index the dice-face}\})$$
$$= \theta_j = \theta_A^{x_A} \times \theta_C^{x_C} \times \theta_G^{x_G} \times \theta_T^{x_T} = \prod_k \theta_k^{x_k} = \theta^x$$

Continuous Distributions

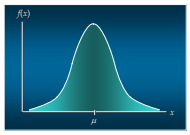
- Uniform Probability Density Function

$$p(x) = 1/(b-a) \quad \text{for } a \le x \le b$$
$$= 0 \quad \text{elsewhere}$$



- Normal (Gaussian) Probability Density Function

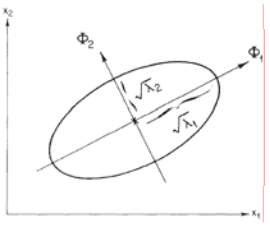
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$



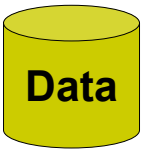
  - The distribution is symmetric, and is often illustrated as a bell-shaped curve.
  - Two parameters, μ (mean) and σ (standard deviation), determine the location and shape of the distribution.
  - The highest point on the normal curve is at the mean, which is also the median and mode.
  - The mean can be any numerical value: negative, zero, or positive.

- Multivariate Gaussian

$$p(X; \vec{\mu}, \Sigma) = \frac{1}{\left(\sqrt{2\pi}\right)^{n/2} |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(X-\vec{\mu})^T \Sigma^{-1} (X-\vec{\mu})\right\}$$



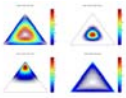
Density Estimation Schemes



$(x_1^{(1)},\ldots,x_n^{(1)})$
$(x_1^{(2)},\ldots,x_n^{(2)})$
…
$(x_1^{(M)},\ldots,x_n^{(M)})$

Learn parameters → Algorithm → Score param

Maximum likelihood	**Analytical**	10^{-5}
Bayesian	**Gradient**	10^{-3}
Conditional likelihood	**EM**	10^{-15}
Margin	**Sampling**	…
…	…	



Parameter Learning from *iid* Data

- Goal: estimate distribution parameters θ from a dataset of N independent, identically distributed (*iid*), fully observed, training cases

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

- 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:

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

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

$$\theta^* = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} \log L(\theta)$$

Example: Bernoulli model

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

  Binary r.v: $x_n = \{0,1\}$
- Model: $P(x) = \begin{cases} 1-\theta & \text{for } x=0 \\ \theta & \text{for } x=1 \end{cases} \quad \Rightarrow \quad P(x) = \theta^x (1-\theta)^{1-x}$
- How to write the likelihood of a single observation x_i?

$$P(x_i) = \theta^{x_i}(1-\theta)^{1-x_i}$$

- The likelihood of dataset $D=\{x_1, \ldots, x_N\}$:

$$P(x_1, x_2, \ldots, x_N \mid \theta) = \prod_{i=1}^{N} P(x_i \mid \theta) = \prod_{i=1}^{N} \left(\theta^{x_i}(1-\theta)^{1-x_i}\right) = \theta^{\sum_{i=1}^{N} x_i}(1-\theta)^{\sum_{i=1}^{N} 1-x_i} = \theta^{\#\text{head}}(1-\theta)^{\#\text{tails}}$$

Maximum Likelihood Estimation

- Objective function:

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

- We need to maximize this w.r.t. θ
- Take derivatives wrt θ

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

Frequency as sample mean

- Sufficient statistics
  - The counts, n_h, where $n_k = \sum_i x_i$, are **sufficient statistics** of data D

Overfitting

- Recall that for Bernoulli Distribution, we have

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

- What if we tossed too few times so that we saw zero head?

  We have $\hat{\theta}_{ML}^{head} = 0,$ and we will predict that the probability of seeing a head next is zero!!!

- The rescue: *"smoothing"*
  - Where n' is know as the pseudo- (imaginary) count

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

  - But can we make this more formal?

Bayesian Parameter Estimation

- Treat the distribution parameters θ also as a *random variable*
- The *a posteriori* distribution of θ after seem the data is:

$$p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{p(D)} = \frac{p(D \mid \theta)\, p(\theta)}{\int p(D \mid \theta)\, p(\theta) d\theta}$$

This is Bayes Rule

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}$$

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

The prior p(.) encodes our prior knowledge about the domain

Frequentist Parameter Estimation

Two people with different priors $p(\theta)$ will end up with different estimates $p(\theta|D)$.

- Frequentists dislike this "subjectivity".
- Frequentists think of the parameter as a fixed, unknown constant, not a random variable.
- Hence they have to come up with different "objective" **estimators** (ways of computing from data), instead of using Bayes' rule.
  - These estimators have different properties, such as being "unbiased", "minimum variance", etc.
  - The maximum likelihood estimator, is one such estimator.

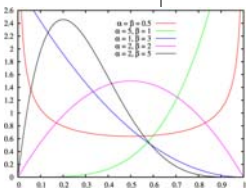
Discussion

θ or $p(\theta)$, this is the problem!

Bayesian estimation for Bernoulli

- Beta distribution:

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



  - When x is discrete $\Gamma(x+1)=x\Gamma(x)=x!$

- 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 isomorphism of the posterior to the prior,
  - such a prior is called a **conjugate prior**
  - α *and* β are hyperparameters (parameters of the prior) and correspond to the number of "virtual" heads/tails (pseudo counts)

Bayesian estimation for Bernoulli, con'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:

Bata parameters can be understood as pseudo-counts

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

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

Effect of Prior Strength

- Suppose we have a uniform prior (α=β=1/2), and we observe $\vec{n} = (n_h = 2, n_t = 8)$
- Weak prior A = 2. Posterior prediction:

$$p(x = h \mid n_h = 2, n_t = 8, \vec{\alpha} = \vec{\alpha}' \times 2) = \frac{1+2}{2+10} = 0.25$$

- Strong prior A = 20. Posterior prediction:

$$p(x = h \mid n_h = 2, n_t = 8, \vec{\alpha} = \vec{\alpha}' \times 20) = \frac{10+2}{20+10} = 0.40$$

- However, if we have enough data, it washes away the prior. e.g., $\vec{n} = (n_h = 200, n_t = 800)$. Then the estimates under weak and strong prior are $\frac{1+200}{2+1000}$ and $\frac{10+200}{20+1000}$, respectively, both of which are close to 0.2

Example 2: Gaussian density

- 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:

$$\ell(\theta; D) = \log P(D \mid \theta) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{n=1}^{N}\frac{(x_n - \mu)^2}{\sigma^2}$$

- MLE: take derivative and set to zero:

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

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

MLE for a multivariate-Gaussian

- It can be shown that the MLE for μ *and* Σ is

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

$$\Sigma_{MLE} = \frac{1}{N}\sum_n (x_n - \mu_{ML})(x_n - \mu_{ML})^T = \frac{1}{N}S$$

$$x_n = \begin{pmatrix} x_n^1 \\ x_n^2 \\ \vdots \\ x_n^K \end{pmatrix} \qquad X = \begin{pmatrix} --- x_1^T --- \\ --- x_2^T --- \\ \vdots \\ --- x_N^T --- \end{pmatrix}$$

where the scatter matrix is

$$S = \sum_n (x_n - \mu_{ML})(x_n - \mu_{ML})^T = \left(\sum_n x_n x_n^T\right) - N\mu_{ML}\mu_{ML}^T$$

- The sufficient statistics are $\Sigma_n x_n$ and $\Sigma_n x_n x_n^T$.
- Note that $X^TX = \Sigma_n x_n x_n^T$ may not be full rank (eg. if $N < D$), in which case Σ_{ML} is not invertible

Bayesian estimation

- Normal Prior:

$$P(\mu) = \left(2\pi\sigma_0^2\right)^{-1/2} \exp\left\{-(\mu - \mu_0)^2 / 2\sigma_0^2\right\}$$

- Joint probability:

$$P(x, \mu) = \left(2\pi\sigma^2\right)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\right\} \times \left(2\pi\sigma_0^2\right)^{-1/2} \exp\left\{-(\mu - \mu_0)^2 / 2\sigma_0^2\right\}$$

- Posterior:

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

where $\tilde{\mu} = \frac{N/\sigma^2}{N/\sigma^2 + 1/\sigma_0^2}\bar{x} + \frac{1/\sigma_0^2}{N/\sigma^2 + 1/\sigma_0^2}\mu_0$, and $\tilde{\sigma}^2 = \left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1}$

Sample mean

Bayesian estimation: unknown μ, known σ

$$\mu_N = \frac{N/\sigma^2}{N/\sigma^2 + 1/\sigma_0^2}\bar{x} + \frac{1/\sigma_0^2}{N/\sigma^2 + 1/\sigma_0^2}\mu_0 , \qquad \tilde{\sigma}^2 = \left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1}$$

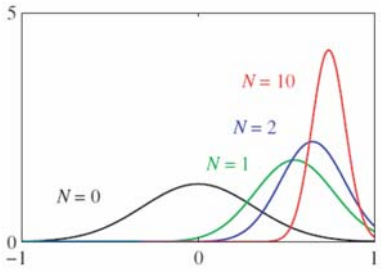
- The posterior mean is a convex combination of the prior and the MLE, with weights proportional to the relative noise levels.
- The precision of the posterior $1/\sigma^2_N$ is the precision of the prior $1/\sigma^2_0$ plus one contribution of data precision $1/\sigma^2$ for each observed data point.

- Sequentially updating the mean
  - $\mu* = 0.8$ (unknown), $(\sigma^2)* = 0.1$ (known)
  - Effect of single data point

    $$\mu_1 = \mu_0 + (x - \mu_0)\frac{\sigma_0^2}{\sigma^2 + \sigma_0^2} = x - (x - \mu_0)\frac{\sigma_0^2}{\sigma^2 + \sigma_0^2}$$

  - Uninformative (vague/ flat) prior, $\sigma^2_0 \to \infty$

    $\mu_N \to \mu_0$



Summary

- Machine Learning is Cool and Useful!!
  - Functional approx.
  - Multivariate statistical models
  - High-dimensional inference
- Learning scenarios:
  - Data
  - Objective function
  - Frequentist and Bayesian
  - Supervised versus unsupervised,
- Density estimation
  - Typical discrete distribution
  - Typical continuous distribution
  - Conjugate priors