Class Registration

- **IF YOU ARE ON THE WAITING LIST**: This class is now fully subscribed. You may want to consider the following options:
  
  - Take the class when it is offered again in the spring semester;
  
  - Come to the first several lectures and see how the course develops. We will admit as many students from the waitlist as we can, once we see how many registered students drop the course during the first two weeks.
Machine Learning 10-701/15-781

- Class webpage:
  - http://www.cs.cmu.edu/~epxing/Class/10701/

Logistics

- Text book
  - Chris Bishop, *Pattern Recognition and Machine Learning* (required)
  - Tom Mitchell, *Machine Learning*
  - David Mackay, *Information Theory, Inference, and Learning Algorithms*

- Mailing Lists:
  - To contact the instructors: 10701-08f-instr@cs.cmu.edu
  - Class announcements list: 10701-08f-announce@cs.cmu.edu.

- TA:
  - Jerry Fu, Doherty Hall 4302E, x8-2039, Office hours: 12-1pm Thu
  - Mark Palatucci, Smith Hall 232.05, x8-2259, Office hours: 1-2pm Wed
  - Suyash Shringarpure, Doherty Hall 4301A, x8-1845, Office hours: 3-4pm Wed
  - Hanghang Tong, Wean Hall 5117, x8-3046, Office hours: 4-5pm Thu

- Class Assistant:
  - Michelle Martin, Wean Hall 4619, x8-5527
  - Diane Stidle, Wean Hall 4612, x8-1299
Logistics

- 5 homework assignments: 30% of grade
  - Theory exercises
  - Implementation exercises
- Final project: 20% of grade
  - Applying machine learning to your research area
    - NLP, IR, Computational biology, vision, robotics …
  - Theoretical and/or algorithmic work
    - A more efficient approximate inference algorithm
    - A new sampling scheme for a non-trivial model …
  - 3-stage reports
- Two exams: 25% of grade each
  - Theory exercises and/or analysis
- Policies …

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

PNAS

© Eric Xing @ CMU, 2006-2008
Learning tasks ...

Speech recognition

Information retrieval

Vision

Games

Evolution

Pedigree

Planning

Machine Learning

Data

Learning Machine

Hypothesis

More, More, More!!

© Eric Xing @ CMU, 2006-2008
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 information, based on the system's own experience with data, and (hopefully) under a explicit 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?

- Speech recognition
- Information retrieval
- Computer vision
- Games
- Robotic control
- Pedigree
- Evolution
- Planning

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!
Object Recognition

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

Robotic Control I

- Now cars can find their own ways!
Robotic Control II

- 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 that …

Text Mining

- We want:
  - Reading, digesting, and categorizing a vast text database is too much for human!
Understanding Brain Activities

Reading a noun (vs verb)

[Reading a noun (vs verb) by Rustandi et al., 2005]
Paradigms of Machine Learning

- **Supervised Learning**
  - Given $D = \{X, Y\}$, learn $f: Y = f(X)$, s.t. $D^{\text{new}} = \{X\} \Rightarrow \{Y\}$

- **Unsupervised Learning**
  - Given $D = \{X\}$, learn $f: Y = f(X)$, s.t. $D^{\text{new}} = \{X\} \Rightarrow \{Y\}$

- **Reinforcement Learning**
  - Given $D = \{\text{env, actions, rewards, simulator/trace/real game}\}$
    
    learn policy: $e, r \rightarrow a$
    utility: $a, e \rightarrow r$
    s.t. $\{\text{env, new real game}\} \Rightarrow a_1, a_2, a_3, \ldots$

- **Active Learning**
  - Given $D = \{X\}$, learn $f: D$ and $D^{\text{new}} \sim G(',)$, s.t. $D'' \Rightarrow G(', policy, \{Y\})$
Theories of Learning

For the learned $F(\theta)$

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

Inference
Prediction
Decision-Making under uncertainty

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

- sickle-cell anemia

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 $\{ <x,y> \}$ of unknown target function $f$

- **Determine:**
  - Hypothesis $h \in H$ that best approximates $f$
Density Estimation

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

![Diagram of Density Estimation]

Basic Probability Concepts

- A **sample space** $S$ is the set of all possible outcomes of a conceptual or physical, repeatable experiment. ($S$ can be finite or infinite.)

  - E.g., $S$ may be the set of all possible outcomes of a dice roll: $S = \{1,2,3,4,5,6\}$

  - E.g., $S$ may be the set of all possible nucleotides of a DNA site: $S = \{A,T,C,G\}$

  - E.g., $S$ may be the set of all possible positions time-space positions of a aircraft on a radar screen: $S = [0, R_{\text{max}}] \times [0, 360^\circ] \times [0, +\infty]$
Random Variable

- A **random variable** is a function that associates a unique numerical value (a token) with every outcome of an experiment. (The value of the r.v. will vary from trial to trial as the experiment is repeated)
  - Discrete r.v.:
    - The outcome of a dice-roll
    - The outcome of reading a nt at site $i$: $X_i$
  - Binary event and indicator variable:
    - Seeing an "A" at a site $\Rightarrow X = 1$, o/w $X = 0$.
    - This describes the true or false outcome a random event.
    - Can we describe richer outcomes in the same way? (i.e., $X=1, 2, 3, 4$, for being A, C, G, T) --- think about what would happen if we take expectation of $X$.
  - Unit-Base Random vector
    - $X = [X_A, X_T, X_G, X_C]'$.
    - $X = [0, 0, 1, 0]'$ ⇒ seeing a "G" at site $i$
  - Continuous r.v.:
    - The outcome of recording the true location of an aircraft: $X_{true}$
    - The outcome of observing the measured location of an aircraft $X_{obs}$

Random Variable

- **Notational convention**
  - Univariate
    - $X_i$: $i \in \{1, \ldots, n\}$
  - Multivariate (random vector)
    - $X = [X_A, X_T, X_G, X_C]'$
    - $X_n = \left[ \begin{array}{c} X_A \\ X_T \\ X_G \\ X_C \end{array} \right]$
Discrete Prob. Distribution

- (In the discrete case), a probability distribution $P$ on $S$ (and hence on the domain of $X$) is an assignment of a non-negative real number $P(s)$ to each $s \in S$ (or each valid value of $X$) such that $\sum_{s \in S} P(s) = 1$. ($0 \leq P(s) \leq 1$)
- Intuitively, $P(s)$ corresponds to the frequency (or the likelihood) of getting $s$ in the experiments, if repeated many times.
- Call $\theta = P(s)$ the parameters in a discrete probability distribution.

- A probability distribution on a sample space is sometimes called a probability model, in particular if several different distributions are under consideration.
- Write models as $M_1$, $M_2$, probabilities as $P(X|M_1)$, $P(X|M_2)$.
- E.g., $M_1$ may be the appropriate prob. dist. if $X$ is from "fair dice", $M_2$ is for the "loaded dice".
- $M$ is usually a two-tuple of (dist. family, dist. parameters).

Discrete Distributions

- Bernoulli distribution: $\text{Ber}(\rho)$
  \[ P(x) = \begin{cases} 
  1 - \rho & \text{if } x = 0 \\
  \rho & \text{if } x = 1 
  \end{cases} \quad \Rightarrow \quad P(X) = \rho^x (1 - \rho)^{1-x} \]

- Multinomial distribution: $\text{Mult}(1, \theta)$
  - Multinomial (indicator) variable:
    \[ X = \begin{bmatrix} X^1 \\ X^2 \\ \vdots \\ X^K \end{bmatrix}, \quad X^i = \{0,1\}, \quad \sum_{i=1}^K X^i = 1 \]
    \[ \text{where } X^i = \text{w.p. } \theta_i, \quad \sum_{i=1}^K \theta_i = 1. \]
  - $p(x(j)) = P(X^j = 1$, where $j$ index the dice - face$)$
    \[ = \theta_j = \theta_1 \times \cdots \times \theta_{j-1} \times \theta_{j+1} \times \cdots \times \theta_K = \prod_{j \neq j} \theta_j = \theta^j \]
Discrete Distributions

- Multinomial distribution: $\text{Mult}(n, \theta)

  - Count variable:

  $$X = \begin{bmatrix} x^1 \\ \vdots \\ x^K \end{bmatrix}, \quad \text{where} \sum_j x^j = n$$

  $$p(x) = \frac{n!}{x^1! \cdots x^K!} \theta_x^n \theta_{x^2} \cdots \theta_{x^K} = \frac{n!}{x^1! \cdots x^K!} \theta^n$$

Density Estimation

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

  - Input Attributes $\rightarrow$ Density Estimator $\rightarrow$ 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, ...)
Density Estimation Schemes

Data

\( (x_1^1, \ldots, x_n^1) \)
\( (x_1^2, \ldots, x_n^2) \)
\( \ldots \)
\( (x_1^K, \ldots, x_n^K) \)

Learn parameters \rightarrow
Algorithm \rightarrow
Score param

Maximum likelihood
Bayesian
Conditional likelihood
Margin

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 \( \theta \) 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:

\[
L(\theta) = \prod_{i=1}^{N} P(x_i; \theta)
\]

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

\[
\hat{\theta} = \arg \max_{\theta} L(\theta) = \arg \max_{\theta} \log L(\theta)
\]
Example: Bernoulli model

- Data:
  - We observed \( N \) iid coin tossings: \( D = \{1, 0, 1, \ldots, 0\} \)

- Representation:
  - Binary r.v.
    \[ x_i \in \{0, 1\} \]

- Model:
  - \[ P(x) = \begin{cases} 1 - \theta & \text{for } x = 0 \\ \theta & \text{for } x = 1 \end{cases} \Rightarrow P(x) = \theta^x (1 - \theta)^{1-x} \]

- 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 \( D = \{x_1, \ldots, x_N\} \):
  - \[ P(x_1, x_2, \ldots, x_N | \theta) = \prod_{i=1}^{N} P(x_i | \theta) = \prod_{i=1}^{N} (\theta^{x_i} (1 - \theta)^{1-x_i}) = \theta^{\sum_{i=1}^{N} x_i} (1 - \theta)^{\sum_{i=1}^{N} (1-x_i)} = \theta^{\text{heads}} (1 - \theta)^{\text{tails}} \]

Maximum Likelihood Estimation

- Objective function:
  - \( \ell(\theta; D) = \log P(D | \theta) = \log \theta^h (1 - \theta)^{n_h} = n_h \log \theta + (N - n_h) \log(1 - \theta) \)

- We need to maximize this w.r.t. \( \theta \)

- Take derivatives wrt \( \theta \)
  - \[ \frac{\partial \ell}{\partial \theta} = \frac{n_h}{\theta} - \frac{N - n_h}{1 - \theta} = 0 \Rightarrow \hat{\theta}_{MLE} = \frac{n_h}{N} \text{ or } \hat{\theta}_{MLE} = \frac{1}{N} \sum x_i \]

- Sufficient statistics
  - The counts, \( n_h \), where \( n_h = \sum x_i \), are sufficient statistics of data \( D \)

© Eric Xing @ CMU, 2006-2008
Overfitting

- Recall that for Bernoulli Distribution, we have
\[
\hat{\theta}_{ML} = \frac{n_{\text{head}}}{n_{\text{head}} + n_{\text{tail}}}
\]

- What if we tossed too few times so that we saw zero head? We have \(\hat{\theta}_{ML} = 0\), 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
\[
\hat{\theta}_{ML} = \frac{n_{\text{head}} + n'}{n_{\text{head}} + n_{\text{tail}} + n'}
\]

- But can we make this more formal?

Bayesian Parameter Estimation

- Treat the distribution parameters \(\theta\) also as a random variable
- The a posteriori distribution of \(\theta\) after seem the data is:
\[
p(\theta | D) = \frac{p(D | \theta) p(\theta)}{p(D)} = \frac{\int p(D | \theta) p(\theta)d\theta}{\int p(D | \theta) p(\theta)d\theta}
\]

This is Bayes Rule

\[\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}\]


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

© Eric Xing @ CMU, 2006-2008
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

$\theta$ or $p(\theta)$, this is the problem!
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 $\theta$:
  \[ P(\theta | x_1, \ldots, x_N) = \frac{P(x_1, \ldots, x_N | \theta) p(\theta)}{p(x_1, \ldots, x_N)} \propto \theta^x (1-\theta)^n \theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1} \]

- Notice the isomorphism of the posterior to the prior,
- such a prior is called a **conjugate prior**
- $\alpha$ and $\beta$ 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 $\theta$:
  \[ P(\theta | x_1, \ldots, x_N) = \frac{p(x_1, \ldots, x_N | \theta) p(\theta)}{p(x_1, \ldots, x_N)} = \theta^n (1-\theta)^{1-n} \propto \theta^{n+\alpha-1} (1-\theta)^{\beta-1} \]

- Maximum a posteriori (MAP) estimation:
  \[ \theta_{MAP} = \arg \max_{\theta} \log P(\theta | x_1, \ldots, x_N) \]

- Posterior mean estimation:
  \[ \theta_{ave} = \int \theta p(\theta | D) d\theta = C \int \theta \theta^{n+\alpha-1} (1-\theta)^{\beta-1} d\theta = \frac{n_0 + \alpha}{N + \alpha + \beta} \]

- Prior strength: $A = \alpha + \beta$
  - $A$ can be interpreted 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 ($\alpha = \beta = 1/2$), and we observe $\bar{n} = (n_0 = 2, n_1 = 8)$
  - Weak prior $A = 2$. Posterior prediction:
    \[ p(x = h | n_0 = 2, n_1 = 8, \alpha = \alpha = 2) = \frac{1+2}{2+10} = 0.25 \]
  - Strong prior $A = 20$. Posterior prediction:
    \[ p(x = h | n_0 = 2, n_1 = 8, \alpha = \alpha = 20) = \frac{10+2}{20+10} = 0.40 \]

- However, if we have enough data, it washes away the prior. e.g., $\bar{n} = (n_0 = 200, n_1 = 800)$. 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
Continuous Prob. Distribution

- A continuous random variable \( X \) can assume any value in an interval on the real line or in a region in a high dimensional space
  - A random vector \( X = [x_1, x_2, \ldots, x_n]^T \) usually corresponds to a real-valued measurements of some property, e.g., length, position, ...

- It is not possible to talk about the probability of the random variable assuming a particular value --- \( P(x) = 0 \)

- Instead, we talk about the probability of the random variable assuming a value within a given interval, or half interval
  - \( P(X \in [a, b]) \)
  - \( P(X < x) = P(X \in [-\infty, x]) \)
  - Arbitrary Boolean combination of basic propositions

Continuous Prob. Distribution

- The probability of the random variable assuming a value within some given interval from \( a \) to \( b \) is defined to be the area under the graph of the probability density function between \( a \) and \( b \).

- Probability mass: \( P(X \in [a, b]) = \int_a^b p(x) \, dx \), note that \( \int_{-\infty}^{\infty} p(x) \, dx = 1 \).

- Cumulative distribution function (CDF):
  \[
P(X = x) = P(X < x) = \int_{-\infty}^{x} p(x') \, dx'
  \]

- Probability density function (PDF):
  \[
p(x) = \frac{d}{dx} P(x)
  \]
  \[
  \int_{-\infty}^{\infty} p(x) \, dx = 1; \quad p(x) > 0, \forall x
  \]
Continuous Distributions

- Uniform Probability Density Function
  \[ p(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{elsewhere} \end{cases} \]

- Normal (Gaussian) Probability Density Function
  \[ p(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]
  - The distribution is symmetric and is often illustrated as a bell-shaped curve.
  - Two parameters, \( \mu \) (mean) and \( \sigma \) (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
  \[ p(X; \mu, \Sigma) = \frac{1}{(\sqrt{2\pi})^{N/2}} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} (X - \mu)^T \Sigma^{-1} (X - \mu) \right\} \]

Example 2: Gaussian density

- Data:
  - We observed \( N \) iid real samples:
    \[ D = \{-0.1, 10, 1, -5.2, \ldots, 3\} \]

- Model:
  \[ P(x) = \left(2\pi\sigma^2\right)^{-1/2} \exp \left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} \]

- Log likelihood:
  \[ \ell(\theta; D) = \log P(D \mid \theta) = -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2} \sum_{x=1}^{N} \frac{(x_{x} - \mu)^2}{\sigma^2} \]

- MLE: take derivative and set to zero:
  \[ \frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2} \sum_{x} (x_{x} - \mu) \quad \Rightarrow \quad \mu_{MLE} = \frac{1}{N} \sum_{x} x_{x} \]
  \[ \frac{\partial \ell}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{x} (x_{x} - \mu)^2 \quad \Rightarrow \quad \sigma_{MLE}^2 = \frac{1}{N} \sum_{x} (x_{x} - \mu_{MLE})^2 \]
MLE for a multivariate-Gaussian

- It can be shown that the MLE for \( \mu \) and \( \Sigma \) is

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

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

where the scatter matrix is

\[
S = \sum_n (x_n - \mu_{MLE})(x_n - \mu_{MLE})^T = (\sum_n x_n x_n^T) - N\mu_{MLE}\mu_{MLE}^T
\]

- The sufficient statistics are \( \Sigma_{x_n} \) and \( \Sigma_{x_n x_n^T} \).

- Note that \( X'X = \Sigma_{x_n x_n^T} \) may not be full rank (eg. if \( N < D \)), in which case \( \Sigma_{MLE} \) 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:

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

where \( \tilde{\mu} = \frac{N/\sigma^2 + 1/\sigma_0^2}{N/\sigma^2 + 1/\sigma_0^2} \mu_0 + \frac{1/\sigma_0^2}{N/\sigma^2 + 1/\sigma_0^2} \mu \), and \( \tilde{\sigma}^2 = \left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1} \).
Bayesian estimation: unknown $\mu$, known $\sigma$

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

- 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
  - $\mu_0 = 0.8$ (unknown), $(\sigma^2)_0 = 0.1$ (known)
  - Effect of single data point
    \[
    \mu_n = \mu_0 + (x - \mu_0) - \frac{\sigma^2_0}{\sigma^2 + \sigma_0^2} = x - (x - \mu_0) - \frac{\sigma^2_0}{\sigma^2 + \sigma_0^2}
    \]
  - Uninformative (vague/flat) prior, $\sigma^2_0 \to \infty$; $\mu_n \to \mu_0$

Summary

- Machine Learning is Cool and Useful!!

- Learning scenarios:
  - Data
  - Objective function
  - Frequentist and Bayesian

- Density estimation
  - Typical discrete distribution
  - Typical continuous distribution (recitation)
  - Conjugate priors

© Eric Xing @ CMU, 2006-2008