Maximum Likelihood Learning

Some material on these is slides borrowed from Andrew Moore's excellent machine learning tutorials located at:

http://www.cs.cmu.edu/~awm/tutorials/

Parameterized Probability Distributions

- Parameterized probability distribution:

$$P(X) = P(X|\theta)$$

- θ - The parameters for the distribution.
- Trivial discrete example: X is a Boolean random variable θ indicates the probability that it will be true.

$\theta = .6$	$p(X = TRUE \mid \theta = .6) = .6$ $p(X = FALSE \mid \theta = .6) = .4$
$\theta = .1$	$p(X = TRUE \mid \theta = .1) = .1$ $p(X = FALSE \mid \theta = .1) = .9$

- θ could also be the mean and covariance of a normal pdf, all of the joint probability tables in a Bayes' net, etc.

Fitting a Distribution to Data

- Assume we have a set of data points x_1 to x_N.
- The goal is to find a distribution that fits that data. I.e. that could have generated the data.
- Two possibilities:
  - Maximum likelihood estimate (MLE) find the parameters that maximize the probability of the data:

$$\hat{\theta} = \underset{\theta}{argmax}\, P(x_{1,} x_{2,} \ldots, x_N | \theta) \quad \text{(we'll do this)}$$

  - Maximum a-priori estimate (MAP) find the parameters that are most probable given the data:

$$\hat{\theta} = \underset{\theta}{argmax}\, P(\theta | x_{1,} x_{2,} \ldots, x_N) \quad \text{(not this)}$$

ML Learning

- We will assume that x_1 to x_N are **iid** – independent and identically distributed.
- So we can rewrite our problem like this (factorization):

$$\hat{\theta} = \underset{\theta}{argmax} \prod_{i=1}^{N} P(x_i|\theta)$$

- Then we can apply our favorite log trick giving us log likelihood:

$$\hat{\theta} = \underset{\theta}{argmax} \sum_{i=1}^{N} \log(P(x_i|\theta)) = \underset{\theta}{argmax}\, LL$$

Maximizing Log Likelihood

- Just another instance of function maximization.
- One approach, set the partial derivatives to 0 and solve:

$$\frac{\partial LL}{\partial \theta_1} = 0$$

$$\frac{\partial LL}{\partial \theta_2} = 0$$

$$\cdots$$

$$\frac{\partial LL}{\partial \theta_K} = 0$$

- If you can't solve it, gradient descent, or your favorite search algorithm.

Silly Example

- Parameterized coin: Theta – probability of heads:
- ***d*** -- vector of toss data, *h* number of heads, *t* number of tails.

$$P(\boldsymbol{d}|\theta) = \prod_{i=1}^{N} P(d_i|\theta) = \theta^h (1-\theta)^t$$

$$L(\boldsymbol{d}|\theta) = log(P(\boldsymbol{d}|\theta)) = h\, log\,\theta + t\, log(1-\theta)$$

$$\frac{\partial L}{\partial \theta} = \frac{h}{\theta} - \frac{t}{1-\theta} = 0 \quad \rightarrow \quad \theta = \frac{h}{h+t}$$

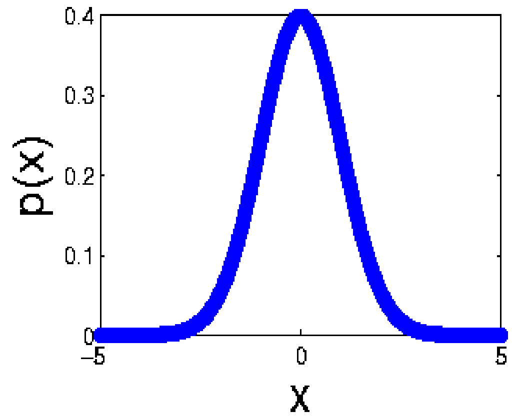
Remember: $\frac{d}{dx} log(x) = 1/x$

Example borrowed from Russel & Norvig

Normal/Gaussian Distribution

- The most useful and oft-seen probability density function in the universe:

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



- σ^2 is the variance, and μ is the mean.

What's So Normal About That?

- The central limit theorem:
  - Assume that X is the sum of N iid random values drawn from some probability distribution.
  - As N increases, the distribution of X approaches the normal distribution.
  - This is true regardless of the distribution of the random values being summed.

Fitting A (scalar) Gaussian Model

$$\theta=(\mu,\sigma) \qquad p(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}} \qquad LL=\sum_{i=1}^{N} log\,\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

Skip a few steps...

$$\frac{\partial LL}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i-\mu) = 0 \qquad \Rightarrow \qquad \mu=\frac{\sum_{i=1}^{N} x_i}{N}$$

$$\frac{\partial LL}{\partial \sigma} = -\frac{N}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{N}(x_i-\mu)^2 = 0 \qquad \Rightarrow \qquad \sigma=\sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$$