**CPSC 550: Machine Learning II** **2008/9 Term 2**

# Lecture 2 — Jan 14, 2009

*Lecturer: Nando de Freitas* *Scribe: Chris Nell*

This lecture introduces fundamental statistical concepts which will be central the rest of the course. The topics presented in these notes are discussed in more detail in the referenced chapters of "All of Statistics" (Wasserman, 2004), freely available for authenticated users of the UBC network at http://www.myilibrary.com/?id=18966.

## Notation

In these notes, we use $F$ to denote a **distribution**, and $f$ a **density**. In particular, we have $F(x) = f(x)dx$, where $dx$ is called the **measure**.

The relationship between these quantities is illustrated for continuous densities and the Lebesgue/Borel measure in Figure 2.1, and for discrete densities using the counting measure (where $F = f$) in Figure 2.2. It is important to note that in the continuous case we measure the probabilities of intervals, not points as we do in the discrete case.

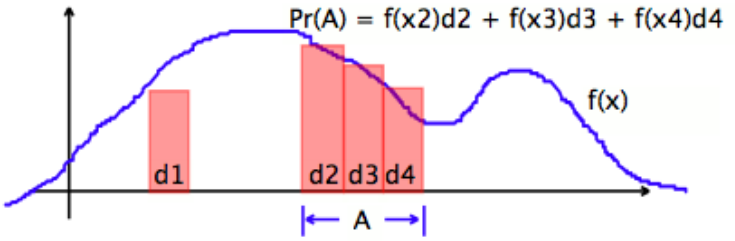


**Figure 2.1.** Example continuous density, distribution, and Lebesgue measure.

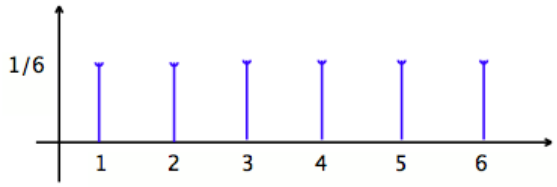


**Figure 2.2.** Example discrete density/distibution. Here, $\Pr(even) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$ which we obtain by counting; hence the name 'counting measure'.

## Chapter 6.1: Introduction to Statistical Learning

The goal of statistical learning is to find the generating distribution $F(\cdot)$ of $X$, given observed data points $X_{1:n} \triangleq \{X_1, X_2, ..., X_n\}$. Example settings for learning include Gaussian processes (Figure 2.3) and mixtures of Gaussians (Figure 2.4).

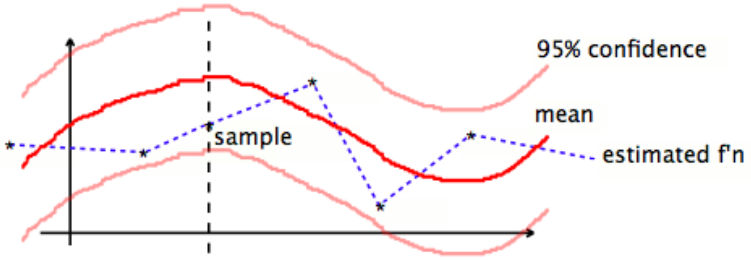


**Figure 2.3.** Learning a Gaussian process. Here, we estimate a function with mean $\mu$ and variance (95% confidence intervals) $\sigma^2$. At each data point, we assume the local distribution (along the vertical line) is Gaussian.

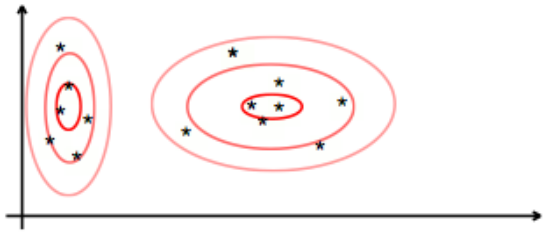

**Figure 2.4.** Learning a mixture of Gaussians. Here, we estimate the parameters of a mixture of two Gaussians, represented via contour map, from observed data points.

## Chapter 6.2: Models

A model is a set of distributions $\mathbb{F}$. We speak of two classes of models:

1. **parametric** models: $\mathbb{F} = \{f(x;\theta), \theta = \Theta\}$
2. **non-parametric** models: $\mathbb{F} = \{$ all distributions $F\}$

When working with parametric models, such as that in Figure 2.5, our goal is to learn the parameters $\theta$ specifying the distribution.

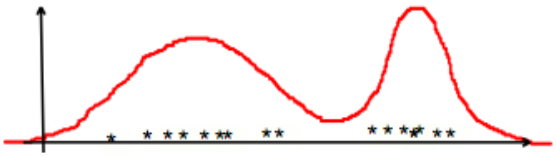

**Figure 2.5.** Mixture of Gaussians: $f(x,\theta) = \pi_1\mathfrak{N}(\mu_1,\sigma_1^2) + \pi_2\mathfrak{N}(\mu_2,\sigma_2^2)$, $\theta = \{\pi_{1:2},\mu_{1:2},\sigma_{1:2}\}$

When working with non-parametric models, our goal is to select the appropriate distribution $F$. The most familiar example might be histogram building (Figure 2.6). Neural networks (Figure 2.7) provide an example where the number of paramters increases rapidly with the number of inputs. Technically, a non-parametric model is a set $\mathbb{F}$ which cannot be characterized by a finite number of parameters.

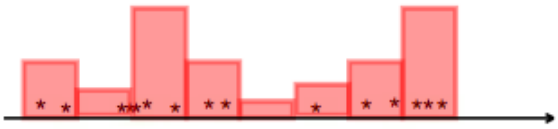

**Figure 2.6.** Histrogramming. While the bin width is an important parameter, it does not characterize the estimated distribution.

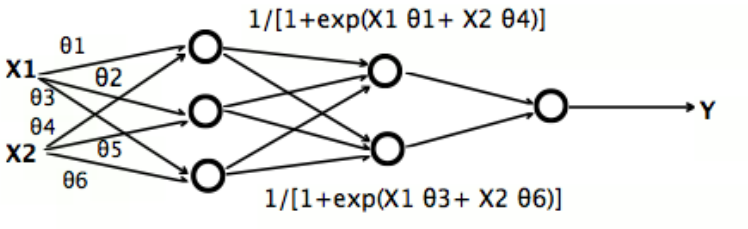


**Figure 2.7.** Neural network. Each node represents a sigmoid function $\frac{1}{1+\exp(\sum \theta_i x_j)}$. The number of parameters $\theta_i$ is exponential in the number of inputs $x_j$.

## Chapter 6.3: Fundamental Concepts

A **point estimate** $\hat{\theta}_n$ of $\theta$ is a "best guess" of $\theta$, based on data $X_{1:n}$:

$$\hat{\theta}_n = g(X_{1:n})$$

A specific example of an estimator is the mean estimator, with $g(X_{1:n}) = \frac{1}{n}\sum_{i=1}^{n} X_i$.

**Bias** is a measure of error due to the inability of an estimtor to account exactly for the observed data, and is defined as:

$$bias(\hat{\theta}_n) = \mathbb{E}_\theta(\hat{\theta}_n) - \theta$$

where $\mathbb{E}_\theta\left(u(x)\right) = \int u(x) f(x;\theta) dx$ is the **expectation** of $u(x)$, and each $x_i \sim F$. Note that while $\hat{\theta}_n$ is a data-dependent random variable, $\theta$ is not; it is a truth.

Examples of high- and low-bias estimators are presented in Figure 2.8. It is clear from the examples that bias alone is not a complete measure of estimator quality. Instead, we seek consistency: $\hat{\theta}_n$ is a consistent estimator of $\theta$ if $\hat{\theta}_n \rightarrow^p \theta$ ($\hat{\theta}_n$ **converges in probability** to $\theta$). Before we can define convergence in probability (next lecture), however, we must introduce some fundamental inequalities.

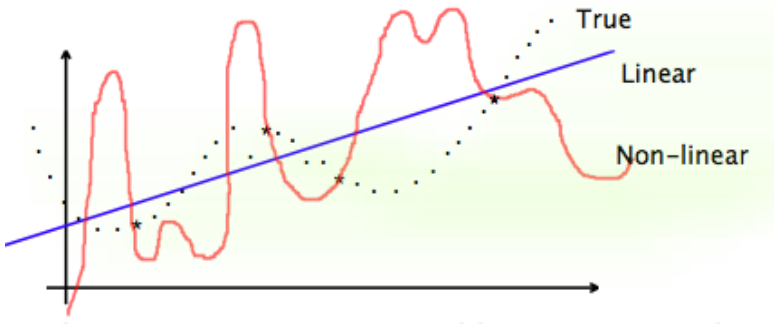


**Figure 2.8.** Bias and variance in estimators. Dotted line represents truth. The linear estimator has high bias but low variance; the non-linear estimator has low bias but high variance.

## Chapter 4.1: Probability Inequalities

The first inequality we will use is **Markov's inequality**:

**Theorem 2.1.** *Let $X$ be a nonnegative random variable; suppose $\mathbb{E}(X)$ exists. For any $t > 0$:*

$$\Pr(X \geq t) \leq \frac{\mathbb{E}(X)}{t}$$

**Proof:** Define the membership **indicator function** for set $S$ as:

$$\mathbb{I}_S(u) = \begin{cases} 1 \text{ if } u \in S \\ 0 \text{ otherwise} \end{cases}$$

The following statement is a tautology over $X \geq 0$ for any $t > 0$:

$$X \geq t\mathbb{I}_{[t,\infty)}(X)$$

Taking expectations:

$$\mathbb{E}(X) \geq t\mathbb{E}\left(\mathbb{I}_{[t,\infty)}(X)\right) = t\Pr\left(X \in [t,\infty)\right) = t\Pr(X \geq t)$$

□

**Proof:** Alternative, adapted from Wasserman. Since $X \geq 0$:

$$\begin{aligned} \mathbb{E}(X) &= \int_0^\infty x f(x)dx = \int_0^t x f(x)dx + \int_t^\infty x f(x)dx \\ &\geq \int_t^\infty x f(x)dx \geq t\int_t^\infty f(x)dx = t\Pr(X \geq t) \end{aligned}$$

□

Next, we introduce **Chebyshev's inequality**:

**Theorem 2.2.** *Let $X$ be a nonnegative random variable; suppose $\mu = \mathbb{E}(X)$ exists, and let $\sigma^2 = \mathbb{V}(X)$, the* **variance** *of $X$. For any $t > 0$:*

$$\Pr(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

**Proof:** Using Markov's inequality, and recalling that $\mathbb{V}(X) = \mathbb{E}(X^2) - \mu^2$:

$$\begin{aligned} \Pr(|X - \mu| \geq t) &= \Pr\left((X-\mu)^2 \geq t^2\right) \\ &\leq \frac{\mathbb{E}\left((X-\mu)^2\right)}{t^2} = \frac{\sigma^2}{t^2} \end{aligned}$$

□