CS649
Sensor Networks
IP Track Lecture 1: Basic Probability and Statistics

I-Jeng Wang
http://hinrg.cs.jhu.edu/wsn06/

Outline

- Why probability and statistics?
- Basic notation and concepts
- Estimation techniques
    - Maximum likelihood estimates
    - Bayes estimates
- Hypothesis testing and detection theory

Why Probability and Statistics for Information Processing in Sensor Networks?

- Significant uncertainties in sensor networks
  - Environment uncertainty (physical environment, unpredictable events, etc.)
  - System uncertainty (network connectivity, node states, power consumptions, etc.)
- Probabilistic and statistical methods provide mathematically consistent tools for modeling and managing uncertainty
- Important to incorporate application-specific domain knowledge in information processing
  - Sensor models (e.g. accuracy of a temperature sensor)
  - Prior information on the underlying physical phenomenon relevant to the application (e.g. air flow models for detection and tracking of potential toxic gas leaks)

Basic concepts and notation

- Probability Model
- Conditional probability and Bayes rule
- Random Variables and Expectation
- Modes of Convergence

Probability Model

- A Sample Space Ω
- A Probability Measure P: a function defined on subsets of Ω such that (Axioms of Probability):
  - $P(\Phi) = 0$, Φ is called the impossible event
  - $P(A) \geq 0$
  - If $A_1, A_2, \ldots$ are events and that are mutually exclusive, then

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$$

  - $P(\Omega) = 1$
- Important consequences of the axioms:

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} P(A_n) \quad if \;\; A_n \subset A_{n+1} \qquad\qquad P\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} P(A_n) \quad if \;\; A_{n+1} \subset A_n$$

Conditional Probability and Bayes Rules

- Conditional Probability (given two events A and B of positive probability)

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

- The Law of Total Probability
  - $B_1, B_2, \ldots$ a sequence of pairwise disjoint events with $\sum_n P(B_n) = 1$
  - For any event A,

$$P(A) = \sum_n P(A \mid B_n) P(B_n)$$

- Bayes Rules

$$\boxed{P(B_k \mid A)} = \frac{P(A \mid B_k)\boxed{P(B_k)}}{\sum_n P(A \mid B_n)\boxed{P(B_n)}}$$

Posterior

Priors

Random Variables

- Random Variable $X : \Omega \to \Re$ such that for any x, the following set is an event $\{\omega : \omega \in \Omega \text{ and } X(\omega) \le x\}$
- Cumulative Distribution Function (cdf)

$$F(x) \equiv P\{X \le x\} \qquad F(x_1, \ldots, x_n) \equiv P\{X_1 \le x_1, \ldots, X_n \le x_n\}$$

- Independence: $X_1, \ldots, X_n$ are independent if

$$F(x_1, \ldots, x_n) \equiv F_{X_1}(x_1) \cdots F_{X_n}(x_n)$$

- Probability Mass and Probability Density Functions

pmf $p(x) = P\{X = x\}, \sum_{\text{all } x} p(x) = 1$ pdf $f(x) = \frac{dF(x)}{dx}, \int_{-\infty}^{\infty} f(x)dx = 1$

- Expectation

$$E(X) = \sum_{\text{all } x} xp(x) \qquad E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

Conditional Distributions and Bayes Rules

- Given two continuous random variables X and Y, the conditional density of X given $Y = y$ is defined as

$$f(x \mid y) = \frac{f(x, y)}{f_Y(y)}, \text{ and}$$

$$F(x \mid y) \equiv P\{X \le x \mid Y = y\} = \int_{-\infty}^{x} f(\tau \mid y) d\tau$$

- Bayes Rules:

$$f(y \mid x) = \frac{f(x \mid y) f_Y(y)}{\int f(x \mid y) f_Y(y) dy}$$

Modes of Convergence

Given a random sequence $\{X_n\}$:

- Almost surely convergence (or convergence with probability one)

$$P\left\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\} = 1$$

- Convergence in probability: for any $\varepsilon > 0$

$$\lim_{n\to\infty} P\left\{\omega : | X_n(\omega) - X_n(\omega) | \geq \varepsilon\right\} = 0$$

- Convergence in distribution: for all x such that $F_X(x)$ is continuous

$$F_{X_n}(x) \to F_X(x)$$

a.s. $\Rightarrow$ i.p. $\Rightarrow$ i.d.

Basic Estimators and Properties

- Problem setting: Given data or observations $X_1,\ldots,X_n$ (typically i.i.d.), estimate a quantity of interest θ (e.g. mean of data)
- Basic properties of an estimator $\hat{\theta}(X_1,\ldots X_n)$
  - Unbiasedness: $$E(\hat{\theta}) = \theta$$
  - Strong Consistency: $$\hat{\theta}_n \xrightarrow{a.s.} \theta$$
- Sample mean and variance: Assume data with the same mean m and variance σ^2

sample mean $$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_n)^2$$ sample variance

  - M_n is unbiased
  - S_n^2 is unbiased if data are uncorrelated
  - Both are strongly consistent if data are i.i.d.

Maximum Likelihood Estimates

- Likelihood Function

$$f(x_1,\ldots,x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta) \quad \text{(if } X_i \text{ independent)}$$

- Maximum Likelihood Estimator (MLE)

$$\hat{\theta}_{MLE} = \arg\max_{\theta\in\Theta} f(x_1,\ldots,x_n \mid \theta)$$

- Basic Properties of MLE
  - Log likelihood function is often used in the maximization
  - Under fairly general conditions, MLE is strongly consistent
  - Do not require distribution information on θ when θ is random
  - May *overfit* the data when n is "small"

MLE and the Least Square Method

- Consider a linear regression with Gaussian noise, we observe $X_i = (x_i, y_i)$, $i = 1,\ldots, n$, where

$$Y_i = a + bx_i + \varepsilon_i, \ \ \varepsilon_i \text{ i.i.d. } N(0, \sigma^2)$$

- The log likelihood function for $\theta = (a, b)$ is proportional to

$$\log(\sigma^{-n}) - \sum_{i=1}^{n} (y_i - ax_i - b)^2 / 2\sigma^2$$

- Hence the MLE is identical to the least square estimate that minimizes the sum of squared errors

$$\sum_{i=1}^{n} (y_i - ax_i - b)^2$$

Bayes Estimates

- Bayes Risks (X: data; $\hat{\theta}(X)$ an estimator of θ)
  - Given Loss function $L(\hat{\theta},\theta)$ and a prior for θ, $\pi(\theta)$,
  - Bayes risk is defined as

$$
\begin{aligned}
L_B(\hat{\theta}) &= \iint L(\hat{\theta},\theta) f(x \mid \theta)\pi(\theta)\,dx\,d\theta \\
&= \int \left[\int L(\hat{\theta},\theta) f(\theta \mid x)\,d\theta \right] f(x)\,dx
\end{aligned}
$$

- Bayes Estimator

posterior loss

$$
\hat{\theta}_B(X) = \arg\max_{\hat{\theta}} L_B(\hat{\theta}) = \arg\max_{\hat{\theta}} \boxed{\int L(\hat{\theta},\theta) f(\theta \mid x)\,d\theta}
$$

- Common Bayes Estimators
  - Quadratic loss $\rightarrow$ Minimum mean-squared error (MMSE) estimator
  - Uniform loss $\rightarrow$ Maximum a posteriori (MAP) estimator

MMSE and MAP

- MMSE
  - Quadratic loss: $L(\hat{\theta}, \theta) = \left\| \hat{\theta} - \theta \right\|^2$
  - MMSE is the conditional mean of θ given X, $\hat{\theta}_{MMSE} = E(\theta \mid X)$
- MAP
  - Uniform loss: $L(\hat{\theta}, \theta) = \begin{cases} 1 & \text{if } \left\| \hat{\theta} - \theta \right\| > \varepsilon \\ 0 & \text{otherwise} \end{cases}$
  - $$\begin{aligned} \hat{\theta}_{MAP} &= \arg\max_{\theta} \underbrace{\boxed{f(\theta \mid X)}}_{\text{posterior density}} \\ &= \arg\max_{\theta} \log f(X \mid \theta) + \log \pi(\theta) - \log f(X) \\ &= \arg\max_{\theta} \log \underbrace{\boxed{f(X \mid \theta)}}_{\text{likelihood}} + \log \underbrace{\boxed{\pi(\theta)}}_{\text{prior}} \end{aligned}$$
  - MAP with uniform prior is equivalent to MLE

Detection and Hypothesis Testing

- Testing of Binary Hypothesis (e.g. change detection)
  - X a random vector with distribution $F_\theta(x)$, $\theta \in \Theta$
  - A binary test (detector) of $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$

$$\phi(x) = \begin{cases} 1 & H_1, x \in R \\ 0 & H_0, x \in A \end{cases}$$

rejection region

acceptance region

- Characteristics of ϕ
  - Probability of false alarm (size): *type I error*

$$P_{FA} = \alpha = \sup_{\theta \in \Theta_0} P_\theta \left\{ \phi(\theta) = 1 \right\} = \sup_{\theta \in \Theta_0} E_\theta [\phi(\theta)]$$

  - Detection probability (power): for each $\theta_1 \in \Theta_1$

$$P_D(\theta_1) = \beta(\theta_1) = P_{\theta_1} \left\{ \phi(\theta) = 1 \right\} = E_{\theta_1} [\phi(\theta)]$$

  - Receiver Operating Characteristics (ROC): P_{FA} versus P_D

The Neyman-Pearson Detectors

- Given a required $P_{FA} = \alpha > 0$, a *likelihood ratio* test of the following form maximizes P_D (most powerful)

$$\phi(x) = \begin{cases} 1, & l(x) > \tau \\ \gamma, & l(x) = \tau \\ 0, & l(x) < \tau \end{cases}$$

$$\boxed{l(x) = \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)}}$$

Likelihood ratio

- Choose the threshold τ and parameter γ to produce the desirable α by solving

$$\alpha = 1 - P_{\theta_0}\{l(X) \le \tau\} + \gamma P_{\theta_0}\{l(X) = \tau\}$$

To Probe Further…

- James L. Johnson, *Probability and Statistics for Computer Science*, Wiley, 2003
- Louis L. Scharf, *Statistical Signal Processing*, Prentice Hall, 2002.