CS 559: Math Review II
Probability Theory

Lecture 2-1

In Jang

ijang@stevens.edu

STEVENS INSTITUTE of TECHNOLOGY

THE INNOVATION UNIVERSITY®

Probability Theory

The Axioms of Probability

- 0 <= P(A) <= 1
- P(True) = 1
- P(False) = 0
- P(A or B) = P(A) + P(B) – P(A and B)

Overview

- Discrete Random Variables
  - Conditional Probability
  - Expected Value
  - Bivariate Distribution
  - Multivariate Distribution
  - Bayesian Probability
- Continuous Random Variable
- The univariate Gaussian
- Maximum-likelihood Estimator

Discrete Random Variables

- Sample space Ω: Possible “states” x of the random variable X (Outcomes of the experiment, output of the system, measurement)
- Either have a finite or countable number of states.
- Events: Possible combination of states (‘subsets of Ω’)

Probability Mass Functions

A function which tells us how likely each possible outcome is:

$$P(X = x) = P_x(x) = P(x)$$

$$\sum_{x \in \Omega} P(x) = 1$$

$$P(A) = P(x \in A) = \sum_{x \in A} P(X = x)$$

Expectation and Variance

- Expectation (or mean):

$$E(x) = \sum_{x} P(X = x)x$$

- Expectation of a function:

$$E\big(f(x)\big) = \sum_{x} P(X = x)f(x)$$

- Moments = expectation of power of *X*:

$$M_k = E(X^k)$$

Expectation and Variance

- Variance: Average (squared) fluctuation from the mean

$$Var(X) = E\left(\left(X - E(X)\right)^2\right) = E(X^2) - E(X)^2$$
$$= M_2 - M_1^2$$

- Standard deviation: square root of variance.

Bivariate Distributions

- Joint Distributions: $P(X = x, Y = y)$, a list of all probabilities of all possible pairs of observations
- Marginal Distribution:

$$P(X = x) = \sum_{y} P(X = x, Y = y)$$

- Conditional Distribution: $P(X = x|Y = y) = P(X = x, Y = y)$
  - $X|Y$ has distribution $P(X|Y)$, where $P(X|Y)$ specifies a “lookup-table” of all possible $P(X = x|Y = y)$

Expectation and Covariance of Bivariate Distributions

- Conditional distributions are just distributions which have a (conditional) mean or variance
- Covariance is the expected value of the product of fluctuations:

$$Cov(X,Y) = E\left(\left(X - E(X)\right)\left(Y - E(Y)\right)\right) = E(XY) - E(X)E(Y)$$
$$Var(X) = Cov(X,X)$$

- One common way to construct bivariate random variables is to have a random variable whose parameter is another random variable.

- Two events are independent if knowing that the first took places tells us nothing about the probability of the second: $P(A|B) = P(A)$
- $P(A)P(B) = P(A \cap B)$
- Two random variables are independent if the joint p.m.f. is the product of the marginals:
  - $P(X = x, Y = y) = P(X = x)P(Y = y)$
- If X and Y are independent, we write $X \perp Y$. Knowing the value of X does not tell us anything about Y.
- If *X* and *Y* are independent, *Cov(X,Y)=0*.
- Mutual information is a measure of how “non-independent" two random variables are.

Multivariate Distributions

- ***X, x*** are vector valued.
- Mean:

$$E(\boldsymbol{X}) = \sum_{x} \boldsymbol{x} P(\boldsymbol{x})$$

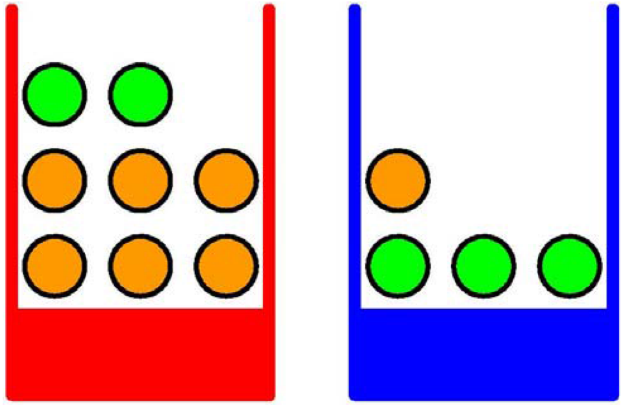
- Covariance Matrix:

$$Cov(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$$
$$Cov(\boldsymbol{X}) = E(\boldsymbol{X}\boldsymbol{X}^T) - E(\boldsymbol{X})E(\boldsymbol{X})^{\boldsymbol{T}}$$

- Conditional and marginal distributions: Can define and calculate any (multi- or single dimensional) marginals or conditional distributions we need: $P(X_1), P(X_1, X_2), P(X_1, X_2, X_3 | X_4)$, etc…

Example

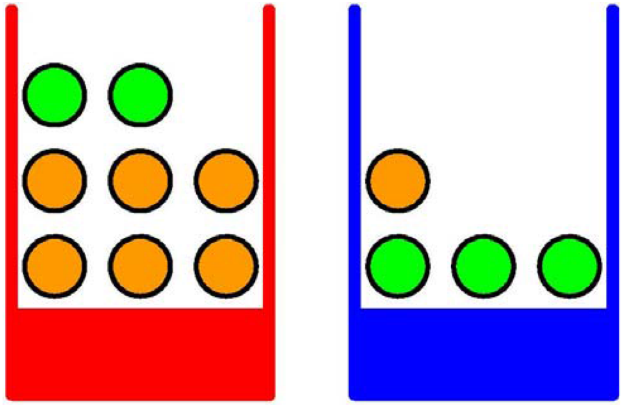
- Let us look at the following example:
  - We have two boxes, one red and one blue
  - Red box: 2 apples and 6 oranges
  - Blue box: 3 apples and 1 orange
  - Pick red box 40% of the time and blue box 60% of the time, then pick one item of fruit



C.M. Bishop, “Pattern Recognition and Machine Learning”, 2006

Example

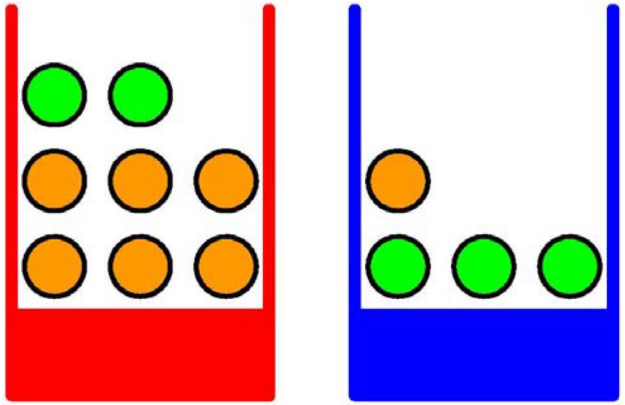
- Define:
  - B random variable for box picked
    - B = {blue(b), red(r)}
  - F identity of fruit
    - F = {apple(a), orange(o)}
  - P(B=r)=0.4 and P(B=b)=0.6
  - Events are mutually exclusive and include all possible outcomes
  - Their probabilities must sum to 1



Example

- $P(B=r) = 0.4$, $P(B=b) = 0.6$
- $P(B=r) + P(B=b) = 1.0$

- Conditional Probabilities
  - $P(F=a|B=r) = 2/8 = 0.25$
  - $P(F=o|B=r) = 6/8 = 0.75$
  - $P(F=a|B=b) = 3/4 = 0.75$
  - $P(F=o|B=b) = 1/4 = 0.25$



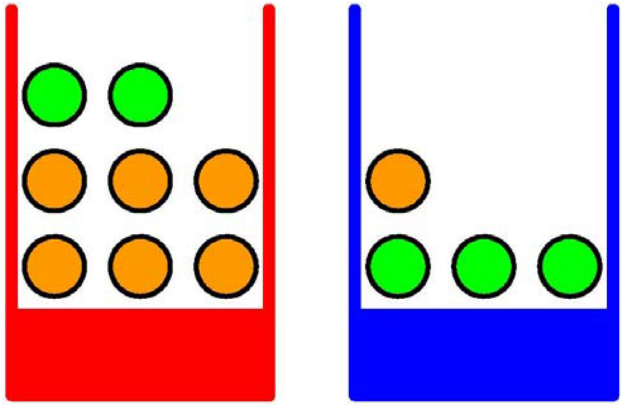
Example

- Note: P(F=a|B=r)+P(F=o|B=r) = 1

- P(F=a) = P(F=a|B=r)P(B=r)+P(F=a|B=b)P(B=b)

  = 1/4 * 4/10 + 3/4 * 6/10 = 11/20

- P(F=o) = 1-11/20=9/20=0.45



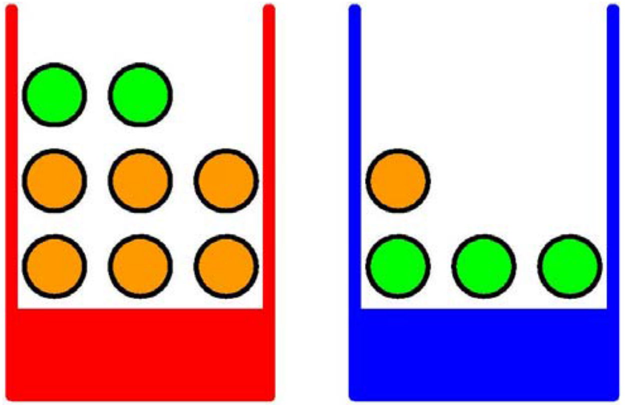
P(F=a|B=r) = 2/8 = 0.25
P(F=o|B=r) = 6/8 = 0.75
P(F=a|B=b) = 3/4 = 0.75
P(F=o|B=b) = 1/4 = 0.25

Bayes Rule on the Fruit Example

- Assume we have now picked an orange
- Ask: which is the probability that it was from the red box?

$$P(B = r|F = o) = \frac{P(F = o|B = r)P(B = r)}{P(F = o)} = \frac{0.75 \times 0.4}{0.45} = \frac{2}{3}$$

P(B=r) = 0.4, P(B=b) = 0.6
P(F=o) = 0.45,
P(F=o|B=r) = 6/8 = 0.75



Prior vs. Posterior

- Prior Probability - If we had been asked which box had been chosen before being told the identity of the selected item of fruit, then the most complete information we have available is provided by the probability P(B).
- Posterior Probability - Once we are told that the fruit is an orange, we can then use Bayes' theorem to compute the probability $P(B|F)$, which we shall call the posterior probability because it is the probability obtained after we have observed F.

Conditional Probability

- Conditional probability: Recalculated probability of event A after someone tells you that event B happened.

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

$$P(A \cap B) = P(A|B)P(B)$$

- Bayes Rule:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Bayesian Probability

- Bayesian view: probabilities provide a quantification of uncertainty. Before observing the data, the assumptions about *w* are captured in the form of a prior probability distribution $P(\boldsymbol{w})$. The effect of the observed data $D = \{(x_1, y_1), \dots, (x_N, y_N)\}$ is expressed by $P(D|\boldsymbol{w})$.
- Bayes' theorem:

$$P(\boldsymbol{w}|D) = \frac{P(D|\boldsymbol{w})P(\boldsymbol{w})}{P(D)}$$

- Bayes' theorem in words: posterior ∝ likelihood × prior

Continuous Random Variables

- A random variable *X* is continuous if its sample space *X* is uncountable.
- In this case, $P(X = x) = 0$ for each *x*.

Continuous Random Variables

- If $p_x(x)$ is a probability density function for *X*, then

$$P(a < X < b) = \int_a^b p(x)dx$$

$$P(a < X < a + dx) \approx p(a) \cdot dx$$

- The cumulative distribution function is $F_x(x) = P(X < x)$. We have that $p_x(x) = F'(x)$, and $F(x) = \int_{-\infty}^{x} p(s)ds$.
- More generally, If A is an event, then

$$P(A) = P(X \in A) = \int_{x \in A} p(x)dx$$

$$P(\Omega) = P(X \in \Omega) = \int_{x \in \Omega} p(x)dx = 1$$

Mean, Variance, and Conditionals

- Mean: $E(x) = \int_x x \cdot p(x) dx$
- Variance: $Var(X) = E(X^2) - E(X)^2$
- If X has pdf *p(x)*, then $X|(X \in A)$ has pdf

$$p_{(x|A)}(x) = \frac{p(x)}{P(A)} = \frac{p(x)}{\int_{x \in A} p(x) dx}$$

- Only makes sense if $P(A) > 0$!

Bivariate Continuous Distributions

- $p_{x,y}(x, y)$, joints probability density function of X and Y
- $\int_x \int_y p(x, y) dx dy = 1$
- Marginal distribution: $p_X(x) = \int_{-\infty}^{\infty} p(x, y) dy$
- Conditional distribution: $p(x|y) = \frac{p(x,y)}{p(y)}$
- Note: $P(Y = y) = 0$!
- Independence: X and Y are independent if $p_{(x,y)}(x, y) = p_x(x) p_y(y)$

The Univariate Gaussian

- Probability density function

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

- Easy to validate:

$$\int_{-\infty}^{\infty} p(x|\mu,\sigma^2)dx = 1$$

- Expectation

$$E(x) = \int_{-\infty}^{\infty} p(x|\mu,\sigma^2)xdx = \mu$$

- Variance

$$Var(x) = E(x^2) - E(x)^2 = \sigma^2$$

Products of Gaussian pdfs

- Suppose $p_1(x) = p\left(x, \mu_1, \frac{1}{\beta_1}\right)$ and $p_2(x) = p\left(x, \mu_2, \frac{1}{\beta_2}\right)$, then

$$p_1(x)p_2(x) \propto p\left(x, \mu, \frac{1}{\beta}\right)$$
$$\beta = \beta_1 + \beta_2$$
$$\mu = \frac{1}{\beta}(\beta_1\mu_1 + \beta_2\mu_2)$$

- In general,

$$p_1(x)p_2(x) \ldots p_n(x) \propto p\left(x, \mu, \frac{1}{\beta}\right)$$
$$\beta = \sum_n \beta_n$$
$$\mu = \frac{1}{\beta}\sum_n \mu_n\beta_n$$

- This is also true for multivariate Gaussians!

ML estimator

- Assuming data points are independent and identically distributed (i.i.d.), the probability of the data set given μ and σ^2 (the likelihood function):

$$P(\boldsymbol{x}|\mu,\sigma^2) = \prod_{n=1}^{N} p(x_n|\mu,\sigma^2)$$

- log-likelihood:

$$\log P(\boldsymbol{x}|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\log\sigma^2 - \frac{N}{2}\log(2\pi)$$

- Maximizing log-likelihood with respect to μ and σ^2:

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n\,, \sigma_{ML}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{ML})^2$$

ML estimator

- The ML solutions μ_{ML} and σ_{ML}^2 are functions of the data set values $x_1, \ldots, x_N$. The expectations of these quantities with respect to the data set values:

$$E(\mu_{ML}) = \mu, E(\sigma_{ML}^2) = \left(\frac{N-1}{N}\right)\sigma^2$$