ICS663: Pattern Recognition

Department of Information and Computer Sciences
University of Hawai`I at Manoa

Kyungim Baek

Announcement

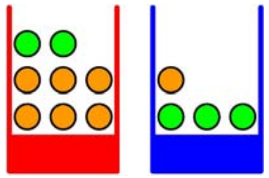
- List of previous projects has been posted

Lecture 2

- Basic Concepts of Probability
- Bayesian Decision Theory
  - Introduction
  - Minimum Probability of Error Classification

Probability Theory

- A simple example: Apples and Oranges



- What are the overall probability that an apple is picked?
- Given that we have chosen an orange, what is the probability that the box we choose was the blue one?

Figure from *PRML* by C. Bishop

Definitions (discrete probability)

- A (stochastic) ***experiment*** is any process that yields one of a given set of possible outcomes and the outcome is not necessarily known in advance
- The ***sample space*** *S* of the experiment is just the set of all possible outcomes
- The ***outcome*** of an experiment is the specific point in a sample space
- An ***event*** *E* is any subset of possible outcomes in *S*, i.e. $E \subseteq S$
- We say that event *E occurs* when the actual outcome *o* is in *E*, which may be written $o \in E$

Probability Theory

- Probability of an event *E* occurring is *P*(*E*):
  - $0 \leq P(E) \leq 1$
  - $\sum_{o \in E} P(o) = 1$
- If *x* is a discrete random variable that can assume any of the values in the finite set $X = \{v_1, v_2, \ldots, v_n\}$, then we denote $P(x = v_k)$ by P_k, the probability of *x* assuming the value v_k
  - $0 \leq P_k \leq 1$
  - $\sum P_k = 1$ for $k = 1, \ldots, n$

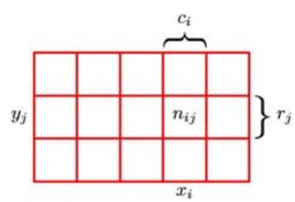
Probability Theory (cont'd)

- Joint Probability

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

- Marginal Probability

$$P(X = x_i) = \frac{c_i}{N}$$



- Conditional Probability

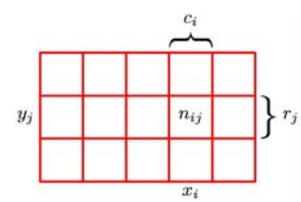
$$P(Y = y_j \mid X = x_i) = \frac{P(Y = y_j, X = x_i)}{P(X = x_i)} = \frac{n_{ij}}{N} \Big/ \frac{c_i}{N} = \frac{n_{ij}}{c_i}$$

Figure from *PRML* by C. Bishop

Rules of Probability

- Sum Rule: $P(X) = \sum_Y P(X, Y)$

$$P(X = x_i) = \frac{c_i}{N} = \frac{1}{N} \sum_j n_{ij}$$
$$= \sum_{y_j \in Y} P(X = x_i, Y = y_j)$$



- Product Rule: $P(X, Y) = P(Y \mid X) P(X)$

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$
$$= \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} = P(Y = y_j \mid X = x_i) P(X = x_i)$$

Figure from *PRML* by C. Bishop

Bayes' Theorem

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

where

$$P(X) = \sum_Y P(X \mid Y)P(Y)$$

Posterior = (Likelihood $\times$ Prior) / Evidence
$\propto$ Likelihood $\times$ Prior

An Example Illustration

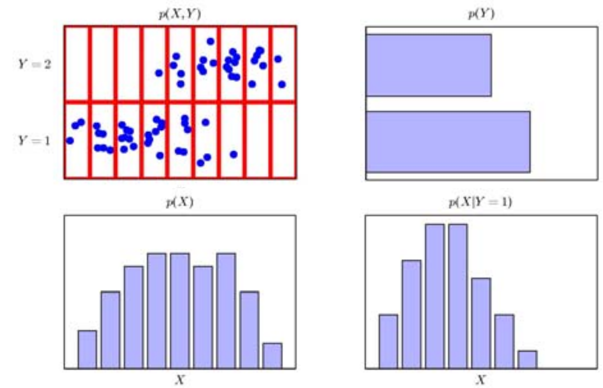
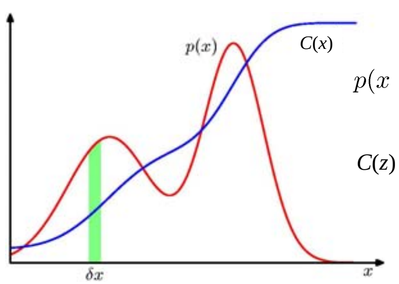


Figure from *PRML* by C. Bishop

Probability Densities



$$p(x \in (a,b)) = \int_a^b p(x)\,\mathrm{d}x$$

$$C(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$

$C(x)$: Cumulative distribution function

$$p(x) \geqslant 0 \qquad \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

From *PRML* by C. Bishop

Expected Value and Variance

- **Expected value**, expectation, mean, or average of a random variable *x*:

$$E(x) = \sum_x xP(x)$$

  – *weighted average* or (*arithmetic*) *mean* of the values of a random variable under the probability distribution
- **Variance** of a random variable *x* :

$$V(x) = E\left[\{x - E(x)\}^2\right] = \sum_x \left[x - E(x)\right]^2 p(x)$$
$$= E(x^2) - \left[E(x)\right]^2$$

Probabilistic Events

- **Independent events** occur independently of each other
  - $P(e_1, e_2, \ldots, e_n) = P(e_1)P(e_2)\cdots P(e_n)$
- **Mutually exclusive events** cannot occur at the same time
  - $P(e_i, e_j) = 0$ for $i \neq j$
  - $P(e_1 \vee e_2 \vee \cdots \vee e_n) = P(e_1) + P(e_2) + \cdots + P(e_n)$
- **Exhaustive events** are a set of events from which at least one of the events will occur
  - $P(e_1 \vee e_2 \vee \cdots \vee e_n) = 1$
- **Mutually exclusive** and **exhaustive events**
  - $P(e_1) + P(e_2) + \cdots + P(e_n) = 1$

Decision Theory

- Decision theory + probability theory
  - Optimal decision making in situations involving uncertainty
- Given (**x**,**t**): input vector & vector of target variables
  - Goal: predict **t** given new value for **x**
  - $p(\mathbf{x},\mathbf{t})$: complete description of the uncertainty
  - Determine $p(\mathbf{x},\mathbf{t})$ from training data set (*inference*)
- Make a specific prediction for **t**, or take a specific action based on the prediction

Bayesian Decision Theory

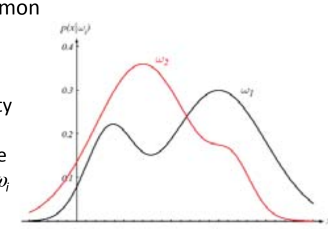
- Based on quantifying the tradeoff between various classification decisions using probability and the costs that accompany such decisions
- The sea bass/salmon example
  - Category (fish type) ω
    - ω is a random variable
    - The catch of salmon and sea bass is equi-probable
      - $P(\omega_1) = P(\omega_2)$ (uniform priors)
      - $P(\omega_1) + P(\omega_2) = 1$ (mutually exclusive and exhaustive)
- *Decision rule* with only the prior information
  - Decide ω_1 if $P(\omega_1) > P(\omega_2)$ otherwise decide ω_2

    $\Rightarrow$ uninformed decision

Informed Decision

- Use of the class-conditional information
  - $p(x|\omega)$: probability that the feature value is x while the category is ω (x is a continuous random variable associated with a pattern feature); usually easier to measure than $P(\omega|x)$
  - $p(x|\omega_1)$ and $p(x|\omega_2)$ describe the difference in lightness between populations of sea bass and salmon

Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i



Recall: Bayes Theorem

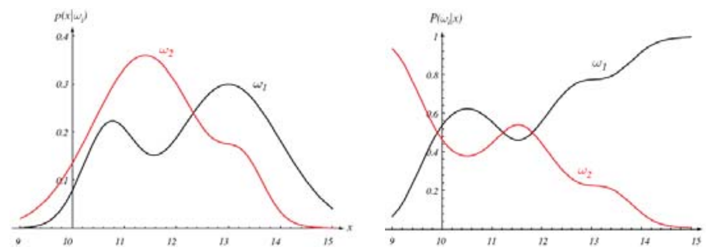
$$P(\omega_j \mid x) = \frac{p(x \mid \omega_j) P(\omega_j)}{p(x)}$$

In case of two categories: $p(x) = \sum_{j=1}^{2} p(x \mid \omega_j) P(\omega_j)$

- $P(\omega_j|x)$: **posterior** probability that the category being ω_j given that the feature value x has been measured (i.e. ω_j conditioned by x)
- $p(x|\omega_j)$: **likelihood** of ω_j with respect to x (i.e. x conditioned by ω_j)
- $P(\omega_j)$: **prior** probability of ω_j occurrence (in the general population)
- $p(x)$: probability of the **evidence** x
- Bayes theorem interpretation:

  posterior = (likelihood × prior) / evidence

Bayes Theorem (cont'd)



Right: Posterior probabilities for the particular priors $P(\omega_1)$ = 2/3 and $P(\omega_2)$ = 1/3 for the class-conditional probability densities on the *left*

Bayesian Classifier

- Decision given the posterior probabilities
  - x is an observation for which:

    if $P(\omega_1|x) > P(\omega_2|x)$ ⟹ True class = ω_1

    if $P(\omega_1|x) < P(\omega_2|x)$ ⟹ True class = ω_2

  Therefore, whenever we observe a particular x, the probability of error is :

    $P(error|x) = P(\omega_1|x)$ if we decide ω_2

    $P(error|x) = P(\omega_2|x)$ if we decide ω_1

- Minimize the average probability of error

$$P(error) = \int_{-\infty}^{\infty} P(error, x)dx = \int_{-\infty}^{\infty} P(error \mid x)\, p(x)dx$$

Bayesian Classifier (cont'd)

- **Bayes decision rule** for minimizing the probability of error

  Decide ω_1 if $P(\omega_1|x) > P(\omega_2|x)$; otherwise decide ω_2

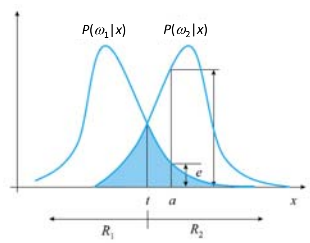
  - $P(error|x) = min\ [P(\omega_1|x), P(\omega_2|x)]$
  - Likelihood $p(x|\omega_j)$: measured during the training
  - Prior probabilities $P(\omega_j)$: assumed to be known within the population of sample patterns
  - Equivalent rule:

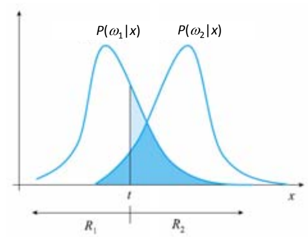
  Decide ω_1 if $p(x|\omega_1) \cdot P(\omega_1) > p(x|\omega_2) \cdot P(\omega_2)$; otherwise decide ω_2

  - The probability of evidence, $p(x)$, is a normalization factor that can be ignored (since it is the same for all class alternatives)

Bayesian Classifier (cont'd)

- Optimal in terms of probability of error
  - Prior probability and the likelihood are assumed to be known
    - ➡ Provide a theoretically optimal classifier





Figures from *PR* by Ilsuk Oh