Problem Set 1

MAS 622J/1.126J: Pattern Recognition and Analysis

Due Monday, 18 September 2006

[Note: All instructions to plot data or write a program should be carried out using either Python accompanied by the `matplotlib` package or Matlab. Feel free to use either or both, but in order to maintain a reasonable level of consistency and simplicity we ask that you do not use other software tools.]

Problem 1: Why?

a. Describe an application of pattern recognition related to your research. What are the features? What is the decision to be made? Speculate on how one might solve the problem. Limit your answer to a page.

b. In the same way, describe an application of pattern recognition you would be interested in pursuing for fun in your life outside of work.

Problem 2: Probability Warm-Up

Let X and Y be random variables. Let $\mu_X \equiv \mathrm{E}[X]$ denote the expected value of X and $\sigma_X^2 \equiv \mathrm{E}[(X-\mu_X)^2]$ denote the variance of X. Use excruciating detail to answer the following:

a. Show $\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y]$.

b. Show $\sigma_X^2 = \mathrm{E}[X^2] - \mu_X^2$.

c. Show that independent implies uncorrelated.

d. Show that uncorrelated does not imply independent.

e. Let $Z = X + Y$. Show that if X and Y are uncorrelated, then $\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2$.

f. Let X_1 and X_2 be independent and identically distributed continuous random variables. Can $\Pr[X_1 \leq X_2]$ be calculated? If so, find its value. If not, explain.

g. Let X_1 and X_2 be independent and identically distributed discrete random variables. Can $\Pr[X_1 \leq X_2]$ be calculated? If so, find its value. If not, explain.

Problem 3: High-Dimensional Probability

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random vector, where the $\{X_i\}$ are independent and identically distributed (i.i.d.) continuous random variables with a uniform probability density function between 0 and 1:

$$p(x_i) = \begin{cases} 1, \text{for } 0 \leq x_i \leq 1 \\ 0, \text{otherwise} \end{cases}$$

Each value $\mathbf{x}$ of the random vector $\mathbf{X}$ can be considered as a point in a n-dimensional hypercube. Since the probability density function of $\mathbf{X}$ is uniform, volume in this n-dimensional space corresponds directly to probability. Find an expression for the percentage of a n-dimensional hypercube's volume located within ϵ of the hypercube's surface. Plot this percentage as a function of n for $1 \leq n \leq 100000$ and $\epsilon = 0.0001$. What do your findings about high-dimensional hypercubes tell you about random variables?

Problem 4: Teatime with Gauss and Bayes

Let $p(x, y) = \frac{1}{2\pi\alpha\beta} e^{-\left(\frac{(y-\mu)^2}{2\alpha^2} + \frac{(x-y)^2}{2\beta^2}\right)}$.

a. Find $p(x)$, $p(y)$, $p(x|y)$, and $p(y|x)$. In addition, give a brief description of each of these distributions.

b. Let $\mu = 0$, $\alpha = 40$, and $\beta = 3$. Plot $p(y)$ and $p(y|x = 13.7)$ for a reasonable range of y. What is the difference between these two distributions?

Problem 5: Covariance Matrix

Let $\Lambda_X = \begin{bmatrix} 64 & -25 \\ -25 & 64 \end{bmatrix}$.

a. Verify that Λ_X is a valid covariance matrix.

b. Find the eigenvalues and eigenvectors of Λ_X by hand. Show all your work.

c. Write a program to find and verify the eigenvalues and eigenvectors of Λ_X.

d. We provide 200 data points sampled from the distribution $\mathcal{N}(0, \Lambda_X)$. Download the dataset from the course website and plot the data points. Project the data onto the covariance matrix eigenvectors and plot the transformed data. What is the difference between the two plots?

Problem 6: Distribution Linearity

Let X_1 and X_2 be i.i.d. according to

$$p(x_i) = \begin{cases} 1, \text{for } 0 \leq x_i \leq 1 \\ 0, \text{otherwise} \end{cases} \quad \text{for } i = 1, 2$$

Let $Y = X_1 + X_2$.

a. Find an expression for $p(y)$. Plot $p(y)$ for some reasonable range of y.

b. Find an expression for $p(x_1|y)$. Plot $p(x_1|y)$ as a function of x_1 with y treated as a known parameter for some reasonable value of y and some reasonable range of x_1.

c. Repeat the parts above, this time letting X_1 and X_2 be i.i.d. according to $\mathcal{N}(0, 1)$.

d. What was the point of this problem? Hint: check out the title.

Problem 7: Monty Hall

To get credit for this problem, you must not only write your own correct solution, but also write a computer simulation (in either Matlab or Python) of the process of playing this game:

Suppose I hide the ring of power in one of three identical boxes while you weren't looking. The other two boxes remain empty. After hiding he ring of power, I ask you to guess which box it's in. I know which box it's in and, after you've made your guess, I deliberately open the lid of an empty box, which is one of the two boxes you did not choose. Thus, the ring of power is either in the box you chose or the remaining closed box you did not choose. Once you have made your initial choice and I've revealed to you an empty box, I then give you the opportunity to change your mind – you can either stick with your original choice, or choose the unopened box. You get to keep the contents of whichever box you finally decide upon.

- What choice should you make in order to maximize your chances of receiving the ring of power? Explain your answer.
- Write a simulation. There are two choices in this game for the contestant in this game: (1) choice of box, (2) choice of whether or not to switch. In your simulation, first let the host choose a random box to place the ring of power. Show a trace of your program's output for a single game play, as well as a cumulative probability of winning for 1000 rounds of the two policies (1) to choose a random box and then switch and (2) to choose a random box and not switch.