ECE 544: Pattern Recognition **Mark Hasegawa-Johnson** **Fall 2013**
Week 1: Intro. and the basics **08/27 - 08/29**

Administrative stuff
Instructor: Mark Hasegawa-Johnson (jhasegaw@illinois.edu)
TA: Sujeeth Bharadwaj (sbhara3@illinois.edu)
Text: Christopher Bishop's *Neural Networks for Pattern Recognition (1995)* or *Pattern Recognition and Machine Learning (2006)*? Vote in class
Office Hours: Wednesdays, 1 - 3 PM 2013 BI (Will be updated once I have a room)
Resources: Course website, compass, email

1 Introduction

Pattern recognition broadly refers to all disciplines (and sub-disciplines) that learn/make predictions from examples. It is best described using the language of probability and statistics, and that is the topic of Week 1. Given some observations of a random variable (or process) X, the idea is to predict another random variable (process), Y. Within a statistical framework, the goal then is to estimate $p(Y|X)$ or the more general $p(X, Y)$ based on model assumptions and whatever data is available. The problem of estimating $p(Y|X)$ or $p(X, Y)$ from data is referred to as **learning/training** the parameters, classifier, etc. The problem of finding the best Y from a given model and a new sample x is called **inference**. Machine learning problems are classified based on the nature of X, Y, and what is available at training and inference.

1.1 Classification

Any problem for which the label Y is known to be discrete is a classification problem. Classification of handwritten digits, image recognition, spam detection are all examples of classification (If you have taken ECE 561, classification is analogous to detection). When a set of examples along with labels $\{(x_i, y_i)\}$ is available for training, we have a **supervised classification** problem; if Y is not available (at training), we have an **unsupervised classification** problem. We will also discuss mixed scenarios such as **semi-supervised** and **unsupervised pretraining**.

1.2 Regression

When Y is continuous, we have to estimate (estimation in ECE 561) its value and these problems are referred to as regression. Predicting the weather (a continuous value) from the time of the day and geographical location, problems in economics (assets, derivatives, etc.) are all good examples of regression.

1.3 Density estimation

In some cases, we only have X available at training, and cannot make any assumptions on Y (whether it's discrete, continuous, or if it even exists). We can only try to estimate $p(X)$, the underlying distribution. Density estimation is therefore in some sense the continuous extension of unsupervised classification/clustering. In the most general setting, no assumptions on $p(X)$ are made and we estimate $p(X)$ directly from the data. We refer to such settings as **non-parametric** (no parameters to estimate).

2 Probability and statistics

In this section, we describe the key ingredients of statistical decision theory. As discussed earlier, let X be an observable random variable, Y, the variable we intend to predict (the label), and $f(X)$ be the prediction for Y. Naturally, we want to test the goodness of some prediction $f(X)$ against the true value, Y. We denote this by $L(Y, f(X))$, a loss function that measures how close the prediction is to the ground-truth label. Given some training data – pairs of (x, y), it is intuitive to minimize the average loss. We call this the **risk** and denote it by

$$R(f) = E[L(Y, f(X))] = \int L(Y, f(X))p(dx, dy) \tag{1}$$

where $p(x, y)$ is the joint density function of X and Y, and $p(dx, dy)$ is shorthand for: $p(x, y)dxdy$ if they are both continuous, and a summation if they are both discrete. Bayes rule, $f_B(X)$ is the predictor that minimizes $R(f)$. That is,

$$f_B = \arg\min_f R(f) \tag{2}$$

2.1 Why is it called Bayes rule and how do we find it?

We can rewrite $R(f)$ as an expectation of conditional expectations:

$$R(f) = E_{XY}[L(Y, f(X)] = E_X E_{Y|X=x}[L(f(X), Y)] \tag{3}$$

Given a particular x, the optimal predictor $f_B(x)$ is one that minimizes the conditional (aka posterior) risk $E_{Y|X=x}[L(f(X), Y)]$. The strategy for finding Bayes optimal predictor, $f_B(x)$ is therefore to:
1) Compute the posterior distribution, $p_{Y|X}(y|x)$ using Bayes rule
2) Compute the posterior risk $E_{Y|X}[L(f(X), Y)] = \int L(f(x), y)p(y|x)dy$.
3) For each x, $f_B(x)$ is that which minimizes the posterior risk

2.2 Least squares is a popular loss function (for regression)

Least squares: $L(y, f(x)) = \|y - f(x)\|_2^2$ (MMSE, most regression problems, etc.). In this case, we have that the posterior risk is $C(a|x) = \int p_{Y|X}(y|x)(a - y)^2 dy$. The first order necessary condition for a minimum is that the derivative of the cost function we wish to minimize ($C(a|x)$ in this case) is 0:

$$0 = \frac{dC(a|y)}{da} = \int p_{Y|X}(y|x)(a - y)dy = a\int p_{Y|X}(y|x)dy - \int p_{Y|X}(y|x)ydy \tag{4}$$

. Hence, the optimal estimate $f_B(x) = \int p_{Y|X}(y|x)ydy = E_{Y|X=x}[Y]$ is the conditional mean.
Homework/Exercise: Show that when the error is the l_1 norm: $L(y, f(x)) = |y - f(x)|$, the *conditional median estimator* is the Bayes estimate.

Linear regression: A special case is when we know structure, for example, that the estimator is linear: $f(x) = x^T\beta$. Plugging this into the loss function results in

$$L(y, f(x)) = \|y - x^T\beta\|_2^2 = (y - x^T\beta)^T(y - x^T\beta) \tag{5}$$

We differentiate with respect to β and set it to 0 (we wish to minimize the loss function) and obtain

$$\beta = (x^Tx)^{-1}x^Ty \tag{6}$$

Polynomial regression: Suppose we want to estimate a sinusoid from 11 sample points, linear regression alone is not sufficient [Bishop example]. A cube is a much better fit, but an 11th degree polynomial "overfits". We will discuss the concept of generalization in later lectures. Note: polynomial regression is also "linear" in nature – all we need to do is "expand" the variables so that the problem is linear. It is the same optimization problem, just over a different set of variables, and a different matrix needs to be inverted.

Homework/Exercise: Relate this to Bayes error and derive conditions under which a linear predictor is Bayes optimal – what has to be true about the distributions of X and Y?

2.3 Minimum probability of error is also popular (for classification)

It does not always make sense to simply consider the squared loss, especially for classification. Suppose we have C classes, $y\epsilon \{0, 1, ..., C-1\}$. The most intuitive loss function is one that directly measures classification error: we assign a cost of 1 whenever $f(x) \neq y$ and 0 otherwise. That is, we pay a fixed price whenever we are wrong. We derive the Bayes optimal rule under this loss. Since Y is now discrete (a classification problem), we are working with summations, not integrals. The posterior risk, $E_{Y|X}[L(f(x), Y)]$ simplifies to:

$$E_{Y|X}[L(f(x), Y)] = \sum_{y=0}^{C-1} L(y, f(x))p_{Y|X}(y|x) \tag{7}$$

and for the Bayes optimal rule, we minimize the risk:

$$f_B(x) = \arg\min_a \sum_{y=0}^{C-1} L(y, a)p_{Y|X}(y|x) = \arg\min_a \sum_{y=0}^{C-1} \mathbf{1}\{y \neq a\}p_{Y|X}(y|x) \tag{8}$$

We can rewrite this sum as $\sum_{y=0}^{C-1} p_{Y|X}(y|x) - p_{Y|X}(a|x) = 1 - p_{Y|X}(a|x)$. Hence,

$$f_B(x) = \arg\min_a 1 - p_{Y|X}(a|x) = \arg\max_a p_{Y|X}(a|x) \tag{9}$$

This is known as the maximum a posteriori (MAP) estimate since this is equiavelent to maximizing $p_{X|Y}(x|y)p(y)$ and not $p_{X|Y}(x|y)$ as is typically done in the maximum likelihood (ML) setting. **Note:** When the prior is uniform (we have no reason to favor one label over the other), the MAP estimate is the same as the ML estimate

Homework/Exercise: This proof is much simpler (in fact, straight from definition) if you can show that $E[\mathbf{1}\{y = f(x)\}] = P[y = f(x)]$, where $\mathbf{1}$ is the indicator function that returns 1 when the event expressed in its argument occurs and 0 otherwise; and $P[.]$ denotes the probability of the event expressed in the argument. Show that $E[\mathbf{1}\{y = f(x)\}] = P[y = f(x)]$ and using this simplification, prove that the Bayes optimal under the 0-1 loss is the MAP estimate.

Let us consider a simpler setting, in which we only have two classes; i.e., $y \epsilon \{0, 1\}$. In some applications (for example, detection of earthquake), it is not sufficient to simply assign a cost of 1 for every mistake. We define two sub-notions of error: missed detection and false alarm. Let 1 denote an earthquake.

false alarm: There is no earthquake, but we incorrectly hypothesize that there is.
missed detection: There is an earthquake, but we miss detecting it and hypothesize that there isn't
Clearly, it is much more disastrous to overlook the occurence of earthquake than it is to incorrectly suggest that there is one.

Homework/Exercise: Think of an application in which a false alarm is worse than a missed detection. Make some assumptions (about the pdfs, pmfs, etc.) to solve it for a particular scenario; compute the Bayes risk under the 0-1 loss. How would it be different for a general loss function parameterized by a and b, in which a is the cost of false alarm, and b is the cost of missed detection.