ECE 901
Lecture 2: Introduction to Classification and Regression

R. Nowak

5/17/2009

1 Pattern Classification

Recall that the goal of classification is to learn a mapping from the feature space, $\mathcal{X}$, to a label space, $\mathcal{Y}$. This mapping, f, is called a *classifier*. For example, we might have

$$\begin{aligned} \mathcal{X} &= \mathbf{R}^d \\ \mathcal{Y} &= \{0,1\}. \end{aligned}$$

We can measure the loss of our classifier using $0-1$ loss; *i.e.*,

$$\ell(\hat{y},y) = \mathbf{1}\{\hat{y} \neq y\} = \left\{ \begin{array}{ll} 1, & \hat{y} \neq y \\ 0, & \hat{y} = y \end{array} \right.$$

Recalling that risk is defined to be the expected value of the loss function, we have

$$R(f) = E_{XY}\left[\ell(f(X),Y)\right] = E_{XY}\left[\mathbf{1}_{\{f(X)\neq Y\}}\right] = P_{XY}\left(f(X) \neq Y\right).$$

The performance of a given classifier can be evaluated by comparing it to the performance of the best possible classification rule, given full knowledge of the problem (in particular the distribution P_{XY}). The performance of such a classifier is known as the Bayes' risk.

Definition 1 (Bayes' Risk) *The Bayes' risk is the infimum of the risk for all classifiers:*

$$R^* = \inf_f R(f).$$

We will prove that the Bayes risk is achieved by the Bayes classifier.

Definition 2 (Bayes Classifier) *The Bayes classifier is the following mapping:*

$$f^*(x) = \left\{ \begin{array}{ll} 1, & \eta(x) \geq 1/2 \\ 0, & otherwise \end{array} \right.$$

where

$$\eta(x) \equiv P_{Y|X}(Y = 1 | X = x).$$

Note that for any x, $f^*(x)$ is the value of $y \in \{0,1\}$ that maximizes $P_{XY}(Y = y | X = x)$, which is the intuitive thing to do.

Theorem 1 (Risk of the Bayes Classifier)

$$R(f^*) = R^*.$$

Proof: Let $g : \mathcal{X} \to \mathcal{Y}$ be any classifier. We will show that $R(g) - R(f^*) \geq 0$. This implies that no classifier performs better than the Bayes classifier. Note that

$$
\begin{aligned}
R(g) - R(f^*) &= P(g(X) \neq Y) - P(f^*(X) \neq Y) \\
&= \int_{\mathcal{X}} P(g(X) \neq Y | X = x) - P(f^*(X) \neq Y | X = x) dP_X(x) \ . \qquad (1)
\end{aligned}
$$

We will show that

$$
P(g(X) \neq Y | X = x) - P(f^*(X) \neq Y | X = x) \geq 0 \ ,
$$

which implies that $R(g) - R(f^*) \geq 0$. For any g,

$$
\begin{aligned}
P(g(X) \neq Y | X = x) &= 1 - P\left(Y = g(X) | X = x\right) \\
&= 1 - [P\left(Y = 1, g(X) = 1 | X = x\right) + P\left(Y = 0, g(X) = 0 | X = x\right)] \\
&= 1 - [E[\mathbf{1}\{Y = 1\}\mathbf{1}\{g(X) = 1\} | X = x] + E[\mathbf{1}\{Y = 0\}\mathbf{1}\{g(X) = 0\} | X = x]] \\
&= 1 - [\mathbf{1}\{g(x) = 1\}E[\mathbf{1}\{Y = 1\} | X = x] + \mathbf{1}\{g(x) = 0\}E[\mathbf{1}\{Y = 0\} | X = x]] \\
&= 1 - [\mathbf{1}\{g(x) = 1\}P\left(Y = 1 | X = x\right) + \mathbf{1}\{g(x) = 0\}P\left(Y = 0 | X = x\right)] \\
&= 1 - [\mathbf{1}\{g(x) = 1\}\eta(x) + \mathbf{1}\{g(x) = 0\}\left(1 - \eta(x)\right)]
\end{aligned}
$$

Next consider the difference

$$
\begin{aligned}
&P\left(g(X) \neq Y | X = x\right) - P\left(f^*(X) \neq Y | X = x\right) \\
&= \eta(x)\left[\mathbf{1}\{f^*(x) = 1\} - \mathbf{1}\{g(x) = 1\}\right] + (1 - \eta(x))\left[\mathbf{1}\{f^*(x) = 0\} - \mathbf{1}\{g(x) = 0\}\right] \\
&= \eta(x)\left[\mathbf{1}\{f^*(x) = 1\} - \mathbf{1}\{g(x) = 1\}\right] - (1 - \eta(x))\left[\mathbf{1}\{f^*(x) = 1\} - \mathbf{1}\{g(x) = 1\}\right] \\
&= \left(2\eta(x) - 1\right)\left(\mathbf{1}\{f^*(x) = 1\} - \mathbf{1}\{g(x) = 1\}\right),
\end{aligned}
$$

where the second equality follows by noting that $\mathbf{1}\{g(x) = 0\} = 1 - \mathbf{1}\{g(x) = 1\}$. Next recall

$$
f^*(x) = \begin{cases} 1, & \eta(x) \geq 1/2 \\ 0, & otherwise \end{cases}
$$

For x such that $\eta(x) \geq 1/2$, we have

$$
\underbrace{(2\eta(x) - 1)}_{\geq 0} \left(\underbrace{\underbrace{\mathbf{1}\{f^*(x) = 1\}}_{1} - \underbrace{\mathbf{1}\{g(x) = 1\}}_{0 \, or \, 1}}_{\geq 0} \right)
$$

and for x such that $\eta(x) < 1/2$, we have

$$
\underbrace{(2\eta(x) - 1)}_{< 0} \left(\underbrace{\underbrace{\mathbf{1}\{f^*(x) = 1\}}_{0} - \underbrace{\mathbf{1}\{g(x) = 1\}}_{0 \, or \, 1}}_{\leq 0} \right),
$$

which implies

$$
(2\eta(x) - 1)\left(\mathbf{1}\{f^*(x) = 1\} - \mathbf{1}\{g(x) = 1\}\right) \geq 0
$$

or

$$
P(g(X) \neq Y | X = x) - P(f^*(X) \neq Y | X = x) \geq 0 \ ,
$$

concluding the proof. $\blacksquare$

Note that while the Bayes classifier achieves the Bayes risk, in practice this classifier is not realizable because we do not know the distribution P_{XY} and so cannot construct $\eta(x)$. It is interesting to write the *excess risk* $R(g) - R^*$ using (1). Note that any classifier $g : \mathcal{X} \to \mathcal{Y}$ is of the form $g(x) = \mathbf{1}\{x \in G\}$ for some set G. Let $G^* = \{x \in \mathcal{X} : \eta(x) \geq 1/2\}$. Now from (1) we have

$$
\begin{aligned}
R(g) - R^* &= \int_{\mathcal{X}} (2\eta(x) - 1) \left(\mathbf{1}\{f^*(x) = 1\} - \mathbf{1}\{g(x) = 1\} \right) dP_X(x) \\
&= \int_{\mathcal{X}} |2\eta(x) - 1| \left| \mathbf{1}\{f^*(x) = 1\} - \mathbf{1}\{g(x) = 1\} \right| dP_X(x) \\
&= \int_{\mathcal{X}} |2\eta(x) - 1| \mathbf{1}\{f^*(x) \neq g(x)\} dP_X(x) \\
&= \int_{G \Delta G^*} |2\eta(x) - 1| dP_X(x) \ ,
\end{aligned}
$$

where $G \Delta G^* = (G \cap \bar{G}^*) \cup (\bar{G} \cap G^*)$ is the symmetric difference between the sets G and G^*, that corresponds to the set of features where the two classifiers disagree upon.

2 Regression

The goal of regression is to learn a mapping from the input space, $\mathcal{X}$, to the output space, $\mathcal{Y}$. This mapping, f, is called a *estimator*. For example, we might have

$$
\begin{aligned}
\mathcal{X} &= \mathbf{R}^d \\
\mathcal{Y} &= \mathbf{R}.
\end{aligned}
$$

We can measure the loss of our estimator using squared error loss; *i.e.,*

$$
\ell(\hat{y}, y) = (y - \hat{y})^2.
$$

Recalling that risk is defined to be the expected value of the loss function, we have

$$
R(f) = E_{XY}[\ell(f(X), Y)] = E_{XY}[(f(X) - Y)^2].
$$

The performance of a given estimator can be evaluated in terms of how close the risk is to the infimum of the risk for all estimator under consideration:

$$
R^* = \inf_f R(f).
$$

Theorem 2 (Minimum Risk under Squared Error Loss (MSE)) *Let* $f^*(x) = E[Y|X = x]$.

$$
R(f^*) = R^*.
$$

Proof: Let $f : \mathcal{X} \to \mathcal{Y}$ be any prediction rule . We have

$$
\begin{aligned}
R(f) &= E\left[(f(X) - Y)^2\right] \\[2mm]
&= E\left[E\left[(f(X) - Y)^2 | X\right]\right] \\[2mm]
&= E\left[E\left[(f(X) - E[Y|X] + E[Y|X] - Y)^2 | X\right]\right] \\[2mm]
&= E[\; \begin{aligned}[t] &E[(f(X) - E[Y|X])^2 | X] \\ &+ 2E\left[(f(X) - E[Y|X])(E[Y|X] - Y)|X\right] \\ &+ E[(E[Y|X] - Y)^2 | X]] \end{aligned} \\[2mm]
&= E[\; \begin{aligned}[t] &E[(f(X) - E[Y|X])^2 | X] \\ &+ 2(f(X) - E[Y|X]) \times 0 \\ &+ E[(E[Y|X] - Y)^2 | X]] \end{aligned} \\[2mm]
&= E\left[(f(X) - E[Y|X])^2\right] + R(f^*).
\end{aligned}
$$

Thus $R(f) \geq R(f^*)$ for any prediction rule f, and therefore $R^* = R(f^*)$. ■

3   Empirical Risk Minimization

Definition 3 (Empirical Risk) *Let $\{X_i, Y_i\}_{i=1}^n \overset{iid}{\sim} P_{XY}$ be a collection of training data. Then the empirical risk is defined as*

$$
\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^n \ell(f(X_i), Y_i).
$$

Empirical risk minimization is the process of choosing a learning rule which minimizes the empirical risk; *i.e.,*

$$
\hat{f}_n = \arg\min_{f \in \mathcal{F}} \hat{R}_n(f).
$$

The main idea behind this approach is that, for a fixed rule f, the empirical risk should be somewhat close to the true risk. In fact the strong law of large numbers says that for a **fixed** rule f

$$
\hat{R}_n(f) \overset{a.s.}{\to} R(f) ,
$$

as $n \to \infty$.

Example 1 (Pattern Classification) *Let $\mathcal{Y} = \{-1, 1\}$ and consider the set of hyperplane classifiers over the feature space $\mathcal{X} = \mathbf{R}^d$ or $[0,1]^d$:*

$$
\mathcal{F} = \left\{ x \mapsto sign(w'x) : w \in \mathbf{R}^d \right\} ,
$$

where $sign(t) = 2\mathbf{1}\{t \geq 0\} - 1$. If we use the notation $f_w(x) \equiv sign(w'x)$, then the set of classifiers can be alternatively represented as

$$
\mathcal{F} = \left\{ f_w : w \in \mathbf{R}^d \right\}.
$$

In this case, the classifier which minimizes the empirical risk is

$$
\begin{aligned}
\hat{f}_n &= \arg\min_{f \in \mathcal{F}} \hat{R}_n(f) \\[2mm]
&= \arg\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{sign(w'X_i) \neq Y_i\}.
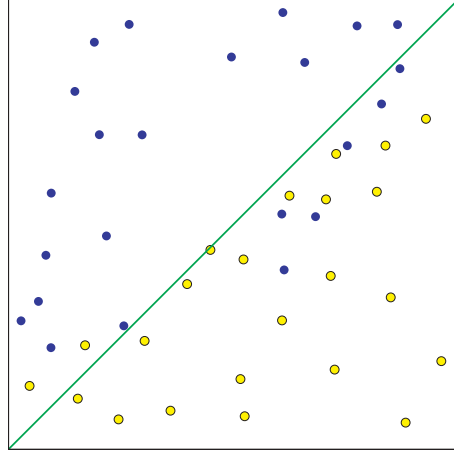\end{aligned}
$$

Figure 1: Example linear classifier for two-class problem.

Example 2 (Regression) *Let the feature space be*

$$\mathcal{X} = [0, 1]$$

and let the set of possible estimators be

$$\mathcal{F} = \{ \text{degree } d \text{ polynomials on } [0, 1] \}.$$

In this case, the classifier which minimizes the empirical risk is

$$
\begin{aligned}
\hat{f}_n &= \arg\min_{f \in \mathcal{F}} \hat{R}_n(f) \\
&= \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2.
\end{aligned}
$$

Alternatively, this can be expressed as

$$
\begin{aligned}
\hat{w} &= \arg\min_{w \in \mathbf{R}^{d+1}} \frac{1}{n} \sum_{i=1}^{n} (w_0 + w_1 X_i + \ldots + w_d X_i^d - Y_i)^2 \\
&= \arg\min_{w \in \mathbf{R}^{d+1}} \|Vw - Y\|^2
\end{aligned}
$$

where V is the Vandermonde matrix

$$
V = \begin{bmatrix}
1 & X_1 & \ldots & X_1^d \\
1 & X_2 & \ldots & X_2^d \\
\vdots & \vdots & \ddots & \vdots \\
1 & X_n & \ldots & X_n^d
\end{bmatrix}.
$$

The pseudoinverse can be used to solve for $\hat{w}$:

$$\hat{w} = (V'V)^{-1}V'Y.$$

A polynomial estimate is displayed in Figure 2.

Note that in some cases, the pseudoinverse of the Vandermonde matrix can produce unstable results. This can be alleviated by using a Chebyshev Vandermonde matrix. While the Vandermonde matrix contains evaluations of the polynomials $\{x^0, x^1, x^2, \ldots, x^k\}$, the Chebyshev Vandermonde matrix contains evaluations of the 0th through k^{th} degree Chebyshev polynomials, which are orthogonal on the interval $[-1, 1]$. See

http://mathworld.wolfram.com/ChebyshevPolynomialoftheFirstKind.html

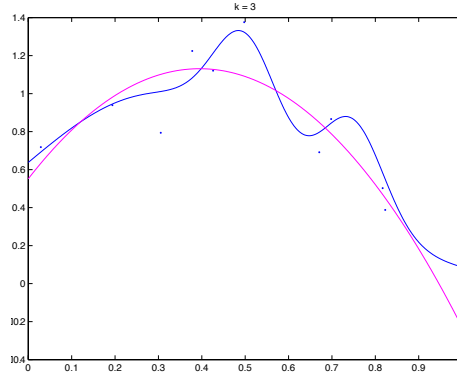for details on the Chebyshev polynomials.

5

Figure 2: Example polynomial estimator. Blue curve denotes f^*, magenta curve is the polynomial fit to the data (denoted by dots).

4   Overfitting

Suppose $\mathcal{F}$, our collection of candidate functions, is very large. We can always make

$$\min_{f \in \mathcal{F}} \hat{R}_n(f)$$

smaller by increasing the cardinality of $\mathcal{F}$, thereby providing more possibilities to fit to the data.

Consider this extreme example: Let $\mathcal{F}$ be all measurable functions. Then every function f for which

$$f(x) = \begin{cases} Y_i, & x = X_i \text{ for } i = 1, \dots, n \\ \text{any value}, & \text{otherwise} \end{cases}.$$

has zero empirical risk $(\hat{R}_n(f) = 0)$. However, clearly this could be a very poor predictor of Y for a new input X .

Example 3 (Classification Overfitting) *Consider the classifier in Figure 3; this demonstrates overfitting in classification. If the data were in fact generated from two Gaussian distributions centered in the upper left and lower right quadrants of the feature space domain, then the optimal estimator would be the linear estimator in Figure 1; the overfitting would result in a higher probability of error for predicting classes of future observations.*

Example 4 (Regression Overfitting) *Below is an m-file that simulates the polynomial fitting. Feel free to play around with it to get an idea of the overfitting problem.*

```
% poly fitting
% rob nowak  1/24/04
clear
close all

% generate and plot "true" function
t = (0:.001:1)';
f = exp(-5*(t-.3).^2)+.5*exp(-100*(t-.5).^2)+.5*exp(-100*(t-.75).^2);
figure(1)
plot(t,f)

% generate n training data & plot
```
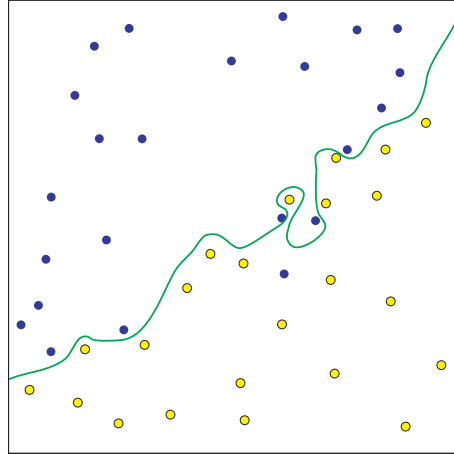
Figure 3: Example of overfitting classifier. The classifier's decision boundary wiggles around in order to correctly label the training data, but the optimal Bayes classifier is a straight line.

```
n = 10;
sig = 0.1; % std of noise
x = .97*rand(n,1)+.01;
y = exp(-5*(x-.3).^2)+.5*exp(-100*(x-.5).^2)+.5*exp(-100*(x-.75).^2)+sig*randn(size(x));
figure(1)
clf
plot(t,f)
hold on
plot(x,y,'.')

% fit with polynomial of order k  (poly degree up to k-1)
k=3;
for i=1:k
    V(:,i) = x.^(i-1);
end
p = inv(V'*V)*V'*y;

for i=1:k
    Vt(:,i) = t.^(i-1);
end
yh = Vt*p;
figure(1)
clf
plot(t,f)
hold on
plot(x,y,'.')
plot(t,yh,'m')
```
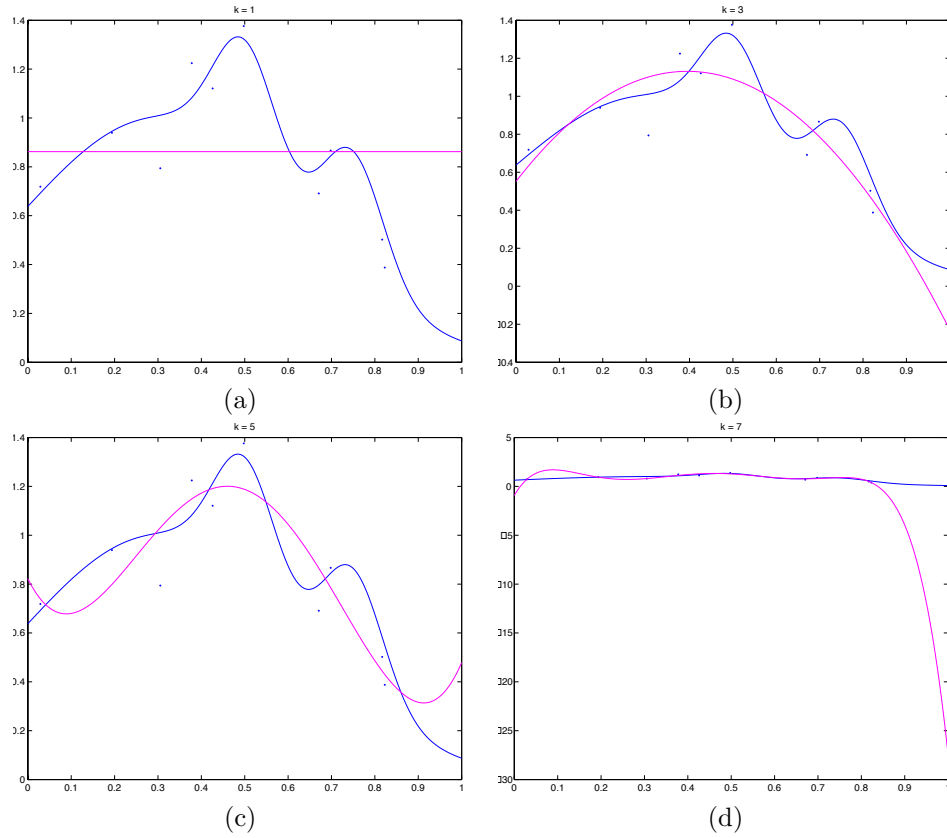
Figure 4: Example polynomial fitting problem. Blue curve is f^*, magenta curve is the polynomial fit to the data (dots). (a) Fitting a polynomial of degree $d = 0$: This is an example of underfitting (b)$d = 2$ (c) $d = 4$ (d) $d = 6$: This is an example of overfitting. The empirical loss is zero, but clearly the estimator would not do a good job of predicting y when x is close to one.