ECE901 Spring 2004 Statistical Regularization and Learning Theory **Lecture: 2**

Statistical Regularization and Learning Theory

Lecturer: Rob Nowak *Scribe: Rebecca Willett*

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.,*

$$\ell(\hat{y},y) = \mathbf{1}_{\{\hat{y}\neq y\}} = \begin{cases} 1, & \hat{y}\neq y \\ 0, & \hat{y}=y \end{cases}$$

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

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

The performance of a given classifier can be evaluated in terms of how close its risk is to the Bayes' risk.

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

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

We can prove that the Bayes' risk is achieved by the Bayes' classifier.

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

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

where

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

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

Theorem 1 (Risk of the Bayes' Classifier)

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

Proof: Let $g(x)$ be any classifier. We will show that

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

For any g,

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

Next note that $\exists g$ such that $\forall x \in \mathbf{R}^d$

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

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

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

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

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

which implies

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

or

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

■

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 know $\eta(x)$.

2 Regression

Recall that the goal of regression is to lear a mapping from the feature space, $\mathcal{X}$, to a label 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 classifier can be evaluated in terms of how close the risk is to the infimum of the risk for all classifiers under consideration:

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

Let $f^*(x) = E_{XY}[Y|X = x]$.

Theorem 2 (Optimal Regression Risk under Squared Error) *For any $f : \mathcal{X} \to \mathcal{Y}$,*

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

Proof:

$$\begin{aligned}
R(f) &= E_{XY}\left[(f(X) - Y)^2\right] \\
&= E_X\left[E_{Y|X}\left[(f(X) - Y)^2|X\right]\right] \\
&= E_X\left[E_{Y|X}\left[(f(X) - E_{XY}[Y|X] + E_{XY}[Y|X] - Y)^2|X\right]\right] \\
&= E_X[\ E_{Y|X}[(f(X) - E_{XY}[Y|X])^2|X] \\
&\qquad +2E_{Y|X}\left[(f(X) - E_{XY}[Y|X])(E_{XY}[Y|X] - Y)|X\right] \\
&\qquad +E_{Y|X}[(E_{XY}[Y|X] - Y)^2|X]] \\
&= E_X[\ E_{Y|X}[(f(X) - E_{XY}[Y|X])^2|X] \\
&\qquad +2(f(X) - E_{XY}[Y|X])E_{Y|X}\left[0\right] \\
&\qquad +E_{Y|X}[(E_{XY}[Y|X] - Y)^2|X]] \\
&= E_{XY}\left[(f(X) - E_{XY}[Y|X])^2\right] + R(f^*).
\end{aligned}$$

Thus if $f^*(x) = E_{XY}[Y|X = x]$, then $R(f^*) = R^*$, as desired. ■

3 Empirical Risk Minimization

Definition 3 (Empirical Risk) *Let $\{X_i, Y_i\}_{i=1}^n$ 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.,*

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

Example 1 (Pattern Classification) *Let the set of possible classifiers be*

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

and let the feature space, $\mathcal{X}$, be $[0, 1]^d$ or $\mathbf{R}^d$. 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

$$\begin{aligned}\hat{f}_n &= \arg\min_{f\in\mathcal{F}} \hat{R}_n(f) \\ &= \arg\min_{w\in\mathbf{R}^d} \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{sign(w'x)\neq y\}}.\end{aligned}$$

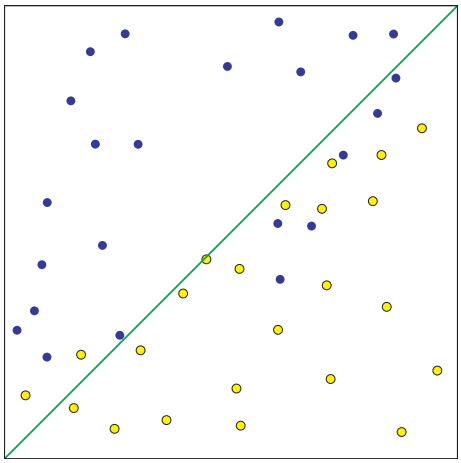

Figure 1: Example linear classifier.

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

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

and let the set of possible estimators be

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

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.

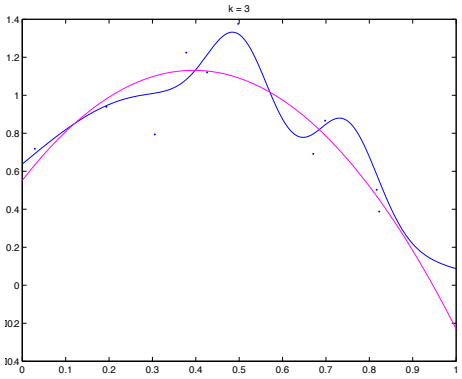

Figure 2: Example polynomial estimator.

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

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

This estimator would have a very small empirical risk ($\hat{R}_n(f) = 0$), but clearly this could be a very poor predictor of Y given a new feature vector 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
```

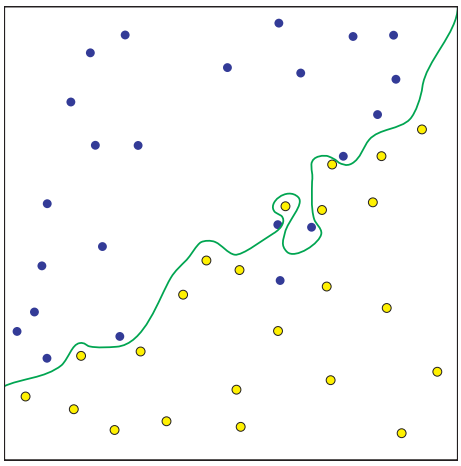

Figure 3: Example of overfitting classifier.

```
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
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')
```

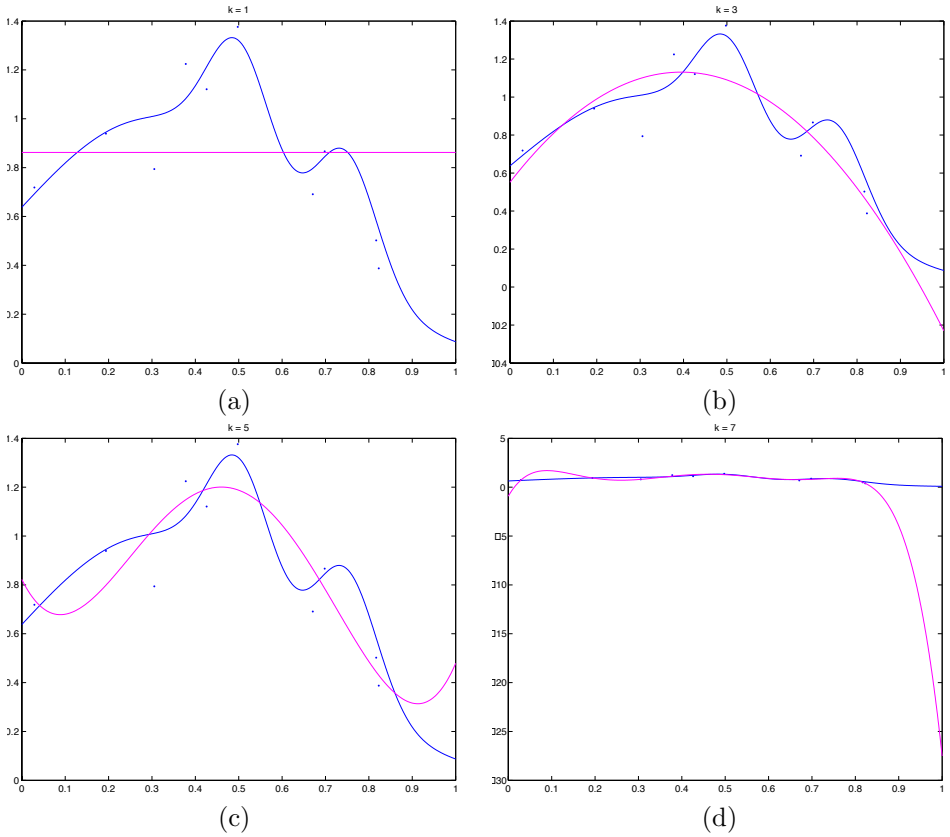

Figure 4: Example polynomial fitting problem. (a) $k = 1$: This is an example of underfitting (b)$k = 3$ (c) $k = 5$ (d) $k = 7$: This is an example of overfitting. The empirical loss is zero, but clearly the estimator would not do a good job of predicting y when x is close to one.