ECE 830 Fall 2010 Statistical Signal Processing

instructor: R. Nowak , **scribe:** N. Sritanyaratana

Lecture 24: Classification

Suppose we want to design an algorithm that will automatically classify objects into one of two categories or "classes". For example, imagine a computer-based system that when presented with an image of a cat or dog must automatically decide which it is using only the image.

This can be done as follows:

1. Locate animal in picture
2. Extract features (e.g., size of ears, length of tail, shape of head, etc.). Collect the features into a vector $x \in \mathbb{R}^d$
3. 3. Provide the computer with a few "labeled" examples; e.g., "this is a picture of a dog". Let the label $Y = 0$ for cats and $Y = 1$ for dogs.

Now given n labeled examples, the computer has pairs

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$$

and will use these to predict the label for unlabelled images. Given an unlabeled image the algorithm extracts the features X and uses the training data $(X_i, Y_i)_i^n = 1$ to predict the correct label for X_j, i.e., to classify the new image as a cat or dog image. How should this be done?

Simple solution: nearest neighbor.

1 Nearest Neighbor (NN) Classifier

1) Find x_i 'closest' to x

2) Assign label associated with closest point in training set

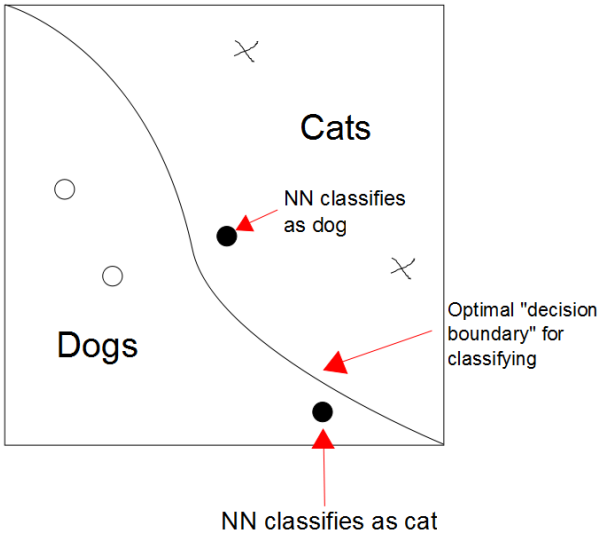
Theorem 1 *Let P_n denote the error rate of the NN classifier based on n samples*

$$\lim_{n\to\infty} P_n \leq 2P_{opt}$$

Notice, P_{opt} is the min error rate possible (Bayes error rate)

The NN classification rule works quite well if n is large, but generally not so well if n is small. The problem is that the N classifier is completely unconstrained; it simply fits as well as possible to the data, and it can "overfit" the data.

Example 1 *Suppose cats and dogs are to be classified using a 2D feature in* $[0,1]^2$.



We will therefore consider systems that aim to determine the best classification rule from a restricted collection of rules.

2 Terminology and Notation

Consider

$(X, Y) \sim P_{xy}$unknown $\qquad\qquad x \in X, y \in 0, 1$

Here, we define X as a feature, Y as a label, and P_{xy} as the unknown distribution relating features and labels.

$(X_i, Y_i)_{i=1}^n \sim P_{xy}$	labeled set of examples; "training data"
$h : X \in 0, 1$	classification rule; "classifier"
$\mathbf{1}_{\{h(x) \neq y\}} = \begin{cases} 0, & h(x) = y \\ 1, & o.w. \end{cases}$	Loss function
$R(h) = \mathbb{E}[\mathbf{1}_{\{h(x) \neq Y\}} = \mathbb{P}(h(x) \neq Y)$	Risk of h (i.e., probability of error or "error rate" of h.)

3 Bayes Classifier

$$h_{opt} = \mathbf{1}_{\{\frac{p(x|y=1)}{p(x|y=0)} > \frac{p(y=0)}{p(y=1)}\}}$$

Notice, $\frac{p(x|y=1)}{p(x|y=0)} > \frac{p(y=0)}{p(y=1)}$ is the likelihood ratio test.

- minimizes probability of error; i.e. Bayes error rate P_{opt}
- defines optimal decision region in X
- impossible to construct without perfect knowledge of class-conditional distributions $p(x|y)$

4 Collections of Classifiers

His a collection of classifiers; i.e., functions $h : X \in \{0, 1\}$

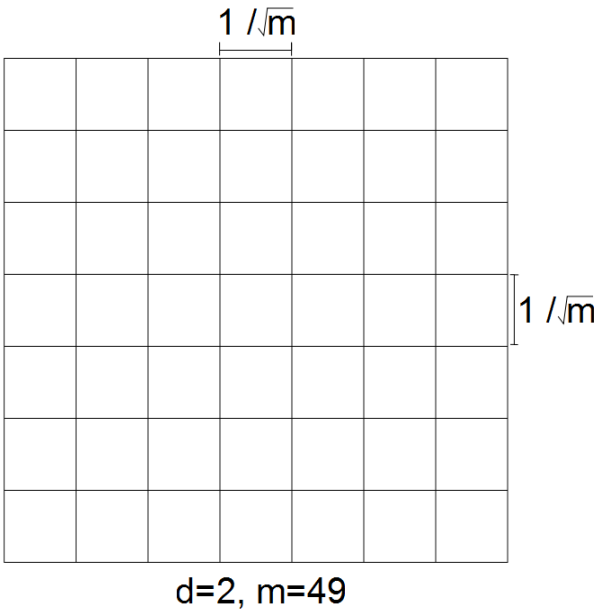
Definition 1 *The* **minimum risk classifier** *in* $\mathbf{H}$*(h^* has min probability of error, but also requires knowledge of P_{xy}) is*

$$h^* = arg \min_{h \in \mathbf{H}} R(h)$$

4.1 Examples of Collections of Classifiers

4.1.1 Histograms

Let $X = [0, 1]^d$ be a partition hypercube into m bins (each with sidelength $m^{-1/2}$)

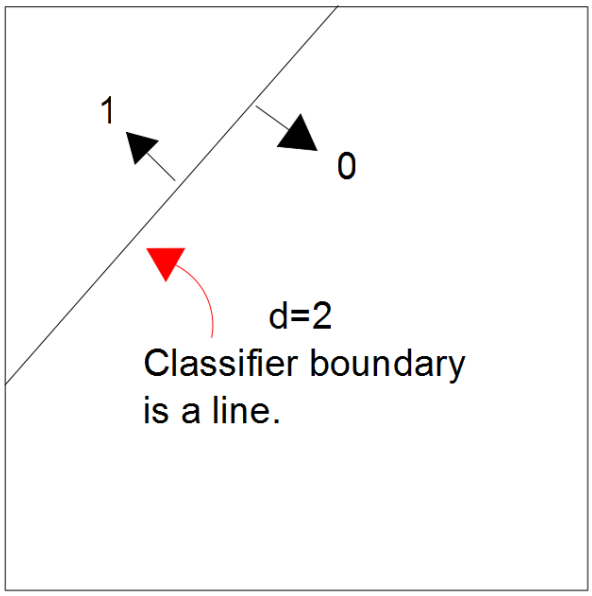


Assign a '1 or '0' label to each bin. There are 2^m possible labelings; each corresponding to a different classifier.

$\mathbf{H} = \{h \in \{\text{one of } 2^m \text{ labelings of m bins}\}\}$ $|\mathbf{H}| = 2^m$. For $h \in \mathbf{H}$, h(x)=label h has on bin that X falls into.

4.1.2 Linear Classifiers

$X = [0,1]^d$ $\mathbf{H} = \{\text{all hyperplanes that split } [0,1]^d \text{ into two halves}\}$ $d = 2$



Fact: if we assume that P_x (distribution of features) satisfies

$$0 < c_0 \leq P(x) \leq c_1, \qquad \forall x$$

then there exists a finite collection of linear classifiers

$$N_\epsilon = \left(\frac{1}{\epsilon}\right)^{d+1}, \qquad \mathbf{H}_\epsilon = \{h_1, \ldots, h_{N_\epsilon}\}$$

such that for any $n \in \mathbf{H}$, $\exists h_i \in \mathbf{H}_\epsilon$ s.t. $|R(h) - R(h_\epsilon)| \leq \epsilon$.

Goal: Use training data $\{(X_i, Y_i)\}$ to select an $h \in \mathbf{H}$ that is (hopefully almost as good as h^*.

5 Empirical Risk Minimization

The "empirical risk" of a classifier h is

$$\hat{R} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{h(x_i) \neq y_i}$$

i.e., the error rate on the training data. Note that

$$\begin{aligned}
\mathbb{E}[\hat{R}(h)] &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\mathbf{1}\{h(x_i) \neq y_i] \\
&= \mathbb{P}(h(x) \neq Y) \qquad \text{since } x_i, y_i \sim P_{xy} \\
&= R(h).
\end{aligned}$$

Since we would like to minimize R(h), this suggests the following procedure for selecting a classifier using the training data:

$$\hat{h} = arg\min_{h \in \mathbf{H}} \hat{R}(h)$$

$\hat{h}$ is called the minimum empirical risk classifier.

6 Analysis of $\hat{h}$

How good is $\hat{h}$? We hope that $R(\hat{h})$ is close to $R(h^*)$. Consider any $h \in \mathbf{H}$.

$$\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{h(x_i)\neq y_i}$$
$$\mathbb{E}[\hat{R}(h)] = R(h)$$

Since the pairs (x_i, y_i) are iid, $\hat{R}(h)$ is an average of bounded iid random variables (in fact, there are iid $Bernoulli(R(h))$.) Therefore we can apply Hoeffding's inequality (also known as the Chernoff bound in the case of sums of Bernoulli r.v.'s) to obtain the bound

$$\mathbb{P}(|R(h) - \hat{R}(h)| > \epsilon) \leq 2e^{-2n\epsilon^2}$$

Approach: Show that

1)$\hat{R}(\hat{h})$ is close to $R(\hat{h})$

2) $R(\hat{h})$ is close to $R(h^*)$

So, can we claim that

$$\mathbb{P}(|R(\hat{h}) - \hat{R}(\hat{h})| > \epsilon) \leq 2e^{-2n\epsilon^2}?$$

No. Since $\hat{h}$ is also a function of the training data, both $R(\hat{h})$ and $\hat{R}(\hat{h})$ are random variables and neither is a simple average of iid variables. But suppose we had a "uniform" bound of the form

$$\mathbb{P}(|R(h) - \hat{R}(h)| > \epsilon) \leq \delta \qquad \forall h \in \mathbf{H}$$

equivalently

$$\mathbb{P}(\max_{h\in\mathbf{H}} |R(h) - \hat{R}(h)| > \epsilon) \leq \delta$$

Then since $\hat{h} \in \mathbf{H}$ it follows that

$$\mathbb{P}(|R(\hat{h}) - \hat{R}(\hat{h})| > \epsilon) \leq \delta$$

So, how do we get a "uniform" bound?

$$\begin{aligned}
&\mathbb{P}(\max_{h} \in \mathbf{H}|R(h) - \hat{R}(h)| > \epsilon) \\
&\qquad = \mathbb{P}(\cup_h \in \mathbf{H}\{|R(h) - \hat{R}(h)| > \epsilon\}) \\
&\qquad \leq \sum_{h\in\mathbf{H}} \mathbb{P}(|R(h) - \hat{R}(h)| > \epsilon) && \text{“union bound”} \\
&\qquad \leq \sum_{h\in\mathbf{H}} 2e^{-2n\epsilon^2} && \text{“Chernoff bound”}
\end{aligned}$$

Now let's assume $\mathbf{H}$ is finite and let $|\mathbf{H}|$ denote the number of classifiers in $\mathbf{H}$. Then

$$\mathbb{P}(\max_{h\in\mathbf{H}} |R(h) - \hat{R})h_| > \epsilon) \leq 2|\mathbf{H}|e^{-2n\epsilon^2} =: \delta$$

Note that in order for this bound to be non-trivial (i.e., $\delta < 1$) we require

$$2|\mathbf{H}|e^{-2n\epsilon^2} < 1$$

$$\implies n > \frac{\log|\mathbf{H}| + \log 2}{2\epsilon^2} = O(\log(|\mathbf{H}|)$$

So now we have that with probability $\leq \delta$

$$\max_h |R(h) - \hat{R}(h)| \geq \epsilon$$

or equivalently, with probability $\geq 1 - \delta$

$$|R(h) - \hat{R}(h)| \leq \epsilon, \forall h \in \mathbf{H}$$

Now we can bound the error rate (risk of $\hat{h}$ as follows. With probability $\geq 1 - \delta$

$$\begin{aligned} R(\hat{h}) &\leq \hat{R}(\hat{h}) + \epsilon, && \text{since } |R(\hat{h}) - \hat{R}(\hat{h})| \leq \epsilon \\ &\leq \hat{R}(h^*) + \epsilon, && \text{since } \min_h \hat{R}(h) = \hat{R}(\hat{h}) \\ &\leq \mathbb{R}(h^*) + 2\epsilon, && \text{since } |R(h^*) - \hat{R}(h^*)| \leq \epsilon \end{aligned}$$

So we have shown that for any $\epsilon > 0$

$$\mathbb{P}(R(\hat{h}) \leq R(h^*) + 2\epsilon) \geq 1 - 2|\mathbf{H}|e^{-2n\epsilon^2}$$

or equivalently for any $\delta \in [0, 1]$

$$\mathbb{P}\left(R(\hat{h}) \leq R(h^*) + 2\sqrt{\frac{\log|H| + \log 2/\delta}{2n}}\right) \geq 1 - \delta$$

i.e., "with high probability $R(\hat{h})$ is not too much larger than $R(h^*)$, provided $n > \log|\mathbf{H}|$"

We can use this probability bound to also bound the expected risk of $\hat{h}$:

$$\mathbb{E}[R(\hat{h})] - R(h^*) \leq 2\sqrt{\frac{\log|H| + \log 2/\delta}{2n}} \cdot (1 - \delta) + 1 \cdot \delta$$
$$2\sqrt{\frac{\log|\mathbf{H} + \log 2/\delta}{2n}} + \delta$$

where $\mathbb{E}[R(\hat{h})]$ is the expectation w.r.t. training data, and $1 \cdot \delta$ is the worst case difference between $R(\hat{h})$ and $R(h^*)$

Since this upper bound is always at least $\frac{1}{\sqrt{n}}$ we may as well set $\delta = \frac{1}{\sqrt{n}}$ to obtain

$$\mathbb{E}[R(\hat{h})] - R(h^*) \leq c_0\sqrt{\log|\mathbf{H}|/n}$$

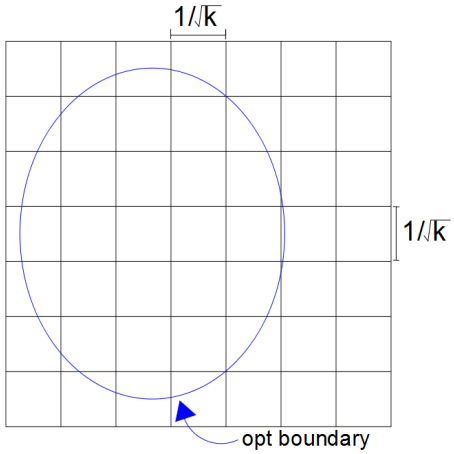
Where $\delta > 0$ is a constant.

7 Histogram Classifiers

$$m \text{ bins} \implies |\mathbf{H}| = 2^m, \log|\mathbf{H}| = m\log 2$$

So for histogram classifiers, on bin j, $\hat{h}$ takes label of "majority" vote of Y_i associated with X_i in bin j.

$$\mathbb{E}[R(\hat{h})] - R(h^*) \leq C_0\sqrt{\frac{m}{n}}$$



where $\frac{m}{n}$ represents the degrees of freedom m divided by n data points.

But what about $R(h^*)$? We can't say much without making some assumption about the optimal (Bayes) classifier h_{opt}. Let $m = k^d$, i.e. each bin has sidelength $1/k$. Furthermore, assume that the boundary of the optimal classifier passes through no more than $C_1 k^{d-1}$ of the bins.

Consider best histogram h^* approximate to h_{opt}. h^* and h_{opt} will agree on all bins except those where boundary of h_{opt} passes through. On those bins the worst case difference is 1. So assuming $p(x) \leq c$ on all bins,

$$R(h^*) - R(h_{opt}) \leq C \cdot \frac{C_1 k^(d-1)}{k^d} = C_2 k^{-1} = C_2 m^{-\frac{1}{d}}$$

So we have

$$\begin{aligned} &\mathbb{E}[R(\hat{h}) - R(h_{opt})] \\ &\qquad = \mathbb{E}[R(\hat{h})] - R(h^*) + R(h^*) - R(h_{opt}) \qquad \leq C_0\sqrt{\frac{m}{n}} + C_2 m^{-\frac{1}{d}} \end{aligned}$$

Choose m to min upper bound $\implies m = n^{\frac{d}{d+2}}$

$$\implies \mathbb{E}[R(\hat{h}) - R(h_{opt}) \leq constant \times n^{-\frac{1}{d+2}}, \qquad \text{Curse of dimensionality}$$

8 Linear Classifiers

Assume $p(x) \geq C > 0, \forall x \in [0,1]^d$
$\mathbf{H}$ is the ϵ-dense set of linear classifiers on $[0,1]^d$. $|\mathbf{H}_\epsilon| = \left(\frac{1}{\epsilon}\right)^{d+1}$

$$\begin{aligned} h_\epsilon^* &= arg\min_{h \in \mathbf{H}_\epsilon} R(h), \qquad R(h_\epsilon^*) - R(h^*) \leq \epsilon \\ h^* &= arg\min_{h \in \mathbf{H}} R(h), \qquad \text{where } \mathbf{H} \text{ is the set of all linear classifiers} \\ \hat{h} &= arg\min_{h \in \mathbf{H}_\epsilon} \hat{R}(h) \end{aligned}$$

$$\mathbb{E}[R(\hat{h})] - R(h_\epsilon^*) \leq C_0 \sqrt{\frac{\log |\mathbf{H}_\epsilon|}{n}}$$
$$= C_0 \sqrt{\frac{(d+1)\log \frac{1}{\epsilon}}{n}}$$

$$\implies \mathbb{E}[R(\hat{h})] - R(h^*) \leq C_0 \sqrt{\frac{(d+1)\log \frac{1}{\epsilon}}{n}} + \epsilon$$

Choose $\epsilon \sqrt{d/n}$

$$\implies \mathbb{E}[R(\hat{h}) - R(h^*)] \leq constant \cdot \sqrt{\frac{d \log n}{n}}$$