ECE 830 Fall 2011 Statistical Signal Processing

instructor: R. Nowak

Lecture 25: 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. 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, i.e., to classify the new image as a cat or dog image. How should this be done? Perhaps the simplest and most intuitive approach is the nearest neighbor classifier.

1 Nearest Neighbor (NN) Classifier

Assume that the features and labels are probabilistically related and have a joint distribution P_{XY}. The NN classifier operates as follows.

1) Find x_i 'closest' to x

2) Assign label associated with closest point in training set

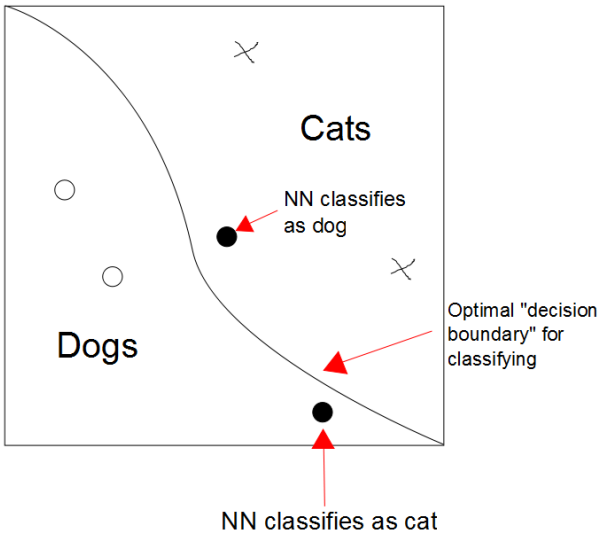
Theorem 1 *Let P_n denote the error rate of the NN classifier based on n samples*

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

P_{opt} is the minimum error rate possible. This is the error rate of the likelihood ratio test based on perfect knowledge of the distribution P_{XY}.

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, as illustrated in the following figure. We will therefore consider systems that aim to determine the best classification rule from a restricted collection of rules.

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



2 Terminology and Notation

Suppose that the features and labels follow a joint distribution

$$(X,Y) \sim P_{XY} .$$

Assume that the features belong to a space $\mathcal{X}$ (e.g., $\mathcal{X} = \mathbb{R}^d$) and the labels are 0 or 1. The binary classification problem involves the following ingredients:

training examples: $(x_i, y_i)_{i=1}^n \overset{iid}{\sim} P_{XY}$

classifiers: $h : \mathcal{X} \to \{0,1\}$, functions mapping features to labels

loss function: $\mathbf{1}_{\{h(x)\neq y\}} = \begin{cases} 0, & h(x) = y \\ 1, & o.w. \end{cases}$

risk: $R(h) = \mathbb{E}[\mathbf{1}_{\{h(X))\neq Y\}}] = \mathbb{P}(h(X) \neq Y)$, where $(X,Y) \sim P_{XY}$

3 Bayes Classifier

If the joint distribution P_{XY} is known, then we can design a classifier that minimizes the risk above. The optimal classifier (called the Bayes Classifier in machine learning) is

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

Notice, $\frac{p(x|y=1)}{p(x|y=0)} > \frac{p(y=0)}{p(y=1)}$ is the likelihood ratio test. The conditional and marginal distributions involved are all known from P_{XY}. A few key points regarding the Bayes Classifier:

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

4 Collections of Classifiers

Let $\mathcal{H}$ denote a collection of classifiers; i.e., functions $h : \mathcal{X} \to \{0, 1\}$

Definition 1 *The* minimum risk classifier *in* $\mathcal{H}$ *is*

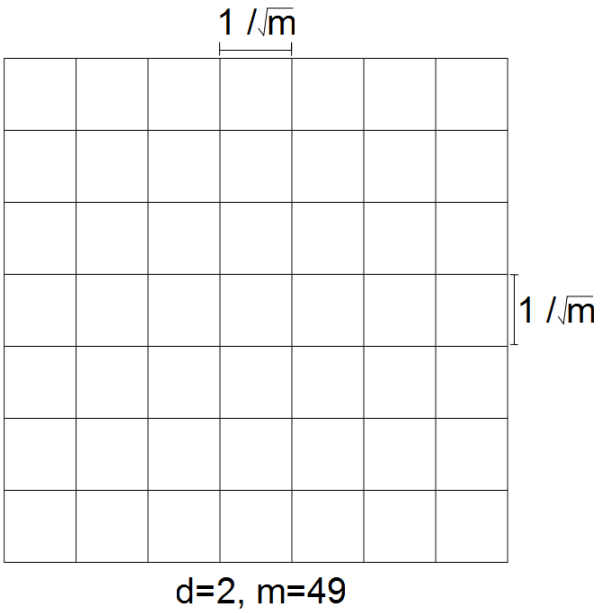
$$h^* = arg \min_{h \in \mathcal{H}} R(h) \ .$$

Finding the mininum risk classifier requires knowledge of the risk R, which is a function of the distribution P_{XY}. So although h^* is usually not easily determined, it serves as a benchmark for comparison.

4.1 Examples of Collections of Classifiers

4.1.1 Histograms

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

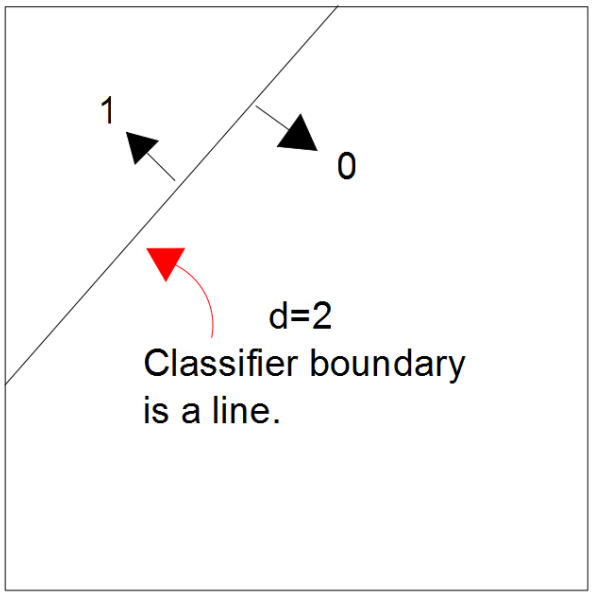


Assign a '1' or '0' label to each bin. There are 2^m possible labelings; each corresponding to a different classifier.

Thus, the number of possible histogram classifiers with m bins is 2^m. We call this the cardinality of the collection $\mathcal{H}$ and denote it by $|\mathcal{H}|$. For any $h \in \mathcal{H}$, a new feature x is assigned the label that h has on the bin that x falls in. **Goal**: Use training data $\{x_i, y_i\}$ to select an $h \in \mathcal{H}$ that is (hopefully almost as good as h^*.

4.1.2 Linear Classifiers

Let $\mathcal{X} = [0, 1]^d$ and consider $\mathcal{H} = \{$all hyperplanes that split $[0, 1]^d$ into two halves$\}$. An example of a linear classifier of this type is shown in the figure below.



If we assume that the marginal density $p(x)$ (density of features) satisfies

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

then there exists a finite collection of linear classifiers

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

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

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

5 Empirical Risk Minimization

The "empirical risk" of a classifier h is

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

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

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

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

$$\widehat{h} \;=\; \arg\min_{h\in\mathcal{H}} \widehat{R}(h)$$

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

6 Analysis of $\widehat{h}$

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

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

Since the pairs (x_i, y_i) are iid, $\widehat{R}(h)$ is an average of bounded iid random variables. In fact, there are iid Bernoulli with probability $R(h)$. 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

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

It is tempting to claim that

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

but this doesn't follow from Hoeffding's inequality since $\widehat{h}$ depends on the training data, and so $\widehat{R}(\widehat{h})$ is *not* a simple average of i.i.d. variables. But suppose we had a "uniform" bound of the form

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

equivalently

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

Then since $\widehat{h} \in \mathcal{H}$ it follows that

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

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

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

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

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

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

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

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

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

$$\max_{h} |R(h) - \widehat{R}(h)| \geq \epsilon$$

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

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

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

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

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

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

or equivalently for any $\delta \in (0,1)$

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

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

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

$$\begin{aligned} \mathbb{E}[R(\widehat{h})] - R(h^*) &\leq 2\sqrt{\frac{\log|H| + \log 2/\delta}{2n}}\,(1-\delta) \;+\; \delta \\ &\leq 2\sqrt{\frac{\log|\mathcal{H} + \log 2/\delta}{2n}} \;+\; \delta \end{aligned}$$

where $\mathbb{E}[R(\widehat{h})]$ is the expectation w.r.t. training data. For this bound we used the fact that 1 is the worst case difference between $R(\widehat{h})$ 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

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

where $c_0 > 0$ is a constant.

7 Histogram Classifiers

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

So for histogram classifiers, on bin j, $\widehat{h}$ takes label of "majority" vote of y_i associated with x_i in bin j.

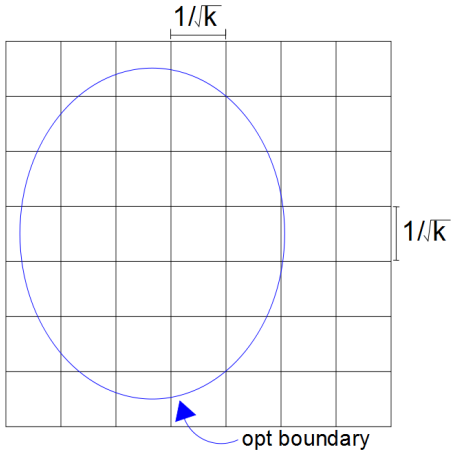
$$\mathbb{E}[R(\widehat{h})] - R(h^*) \leq C_0\sqrt{\frac{m}{n}}$$

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_1k^{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,

$$R(h^*) - R(h_{opt}) \leq C \cdot \frac{C_1k^(d-1)}{k^d} = C_2k^{-1} = C_2m^{-\frac{1}{d}}$$



So we have

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

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

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

8 Linear Classifiers

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

$$h_\epsilon^* = arg\min_{h\in\mathcal{H}_\epsilon} R(h), \qquad R(h_\epsilon^*) - R(h^*) \leq \epsilon$$
$$h^* = arg\min_{h\in\mathcal{H}} R(h), \qquad \text{where H is the set of all linear classifiers}$$
$$\widehat{h} = arg\min_{h\in\mathcal{H}_\epsilon} \widehat{R}(h)$$

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

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

Choose ϵ $\sqrt{d/n}$

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