Learning Patterns

Carlos C. Rodríguez
http://omega.albany.edu:8008/

September 19, 2004

Set up

We assume labeled data $X \in R^d$ with label $Y \in \{0, 1\}$. For example $X = x$ could be the vector of pixels of a digital image of a human face and $Y = 1$ may mean that x is the face of a woman instead of a man. The random vector (X, Y) has distribution specified by, for example, a probability measure μ for X and the regression function $\eta(x)$ of Y on X. We have,

$$P\{X \in A\} = \int_A \mu(dx)$$

and

$$\eta(x) = P\{Y = 1|X = x\}$$

Clearly, by the sum and product rules of probability, we have

$$P\{(X, Y) \in C\} = \int_{C_0} (1 - \eta(x))\mu(dx) + \int_{C_1} \eta(x)\mu(dx)$$

where C_0 contains all the x's such that $(x, 0) \in C$. The joint distribution of (X, Y) can be specified in other ways. Instead of picking first x according to μ and then flipping a coin with probability $\eta(x)$ to determine the label, we could also do it backwards.

First choose a label with (prior) probability $P\{Y = 1\} = \pi$ and then choose x with density (assuming there is one) f_0 or f_1 depending on the previously generated label y. Hence, for $y \in \{0, 1\}$

$$P\{X \in A|Y = y\} = \int_A f_y(x)dx$$

By conditioning of Y we obtain,

$$P\{X \in A\} = (1 - \pi) \int_A f_0(x)dx + \pi \int_A f_1(x)dx = \int_A \mu(dx)$$

Thus, if μ has a density $f(x) = \mu(dx)/dx$ it must be a mixture,

$$f(x) = (1 - \pi) f_0(x) + \pi f_1(x)$$

Also, by the holy theorem,

$$\eta(x) = P\{Y = 1 | X = x\} = \frac{f_1(x)\pi}{f(x)}$$

The Bayes Classifier

A classifier is a map $\delta : R^d \to \{0, 1\}$ that assigns $0 - 1$ labels to data vectors $x \in R^d$. The quality of a classifier δ is measured by its risk $R(\delta)$ which is simply its probability of error,

$$R(\delta) = P\{\delta(X) \neq Y\}$$

The Bayes rule is the classifier δ^* with smallest possible risk, so that for all δ,

$$R(\delta^*) \leq R(\delta)$$

We call $R^* = R(\delta^*)$ the Bayes risk. It provides a measure of difficulty for the pattern recognition task at hand. It is a characteristic of the joint distribution of (X, Y).

A standard result of statistical decision theory is that the Bayes rule for the all-nothing (i.e. 0-1) loss function is the mode of the posterior distribution. Thus, $\delta^*(x) = 1[\eta(x) > 1/2]$.

Here is an easy proof for the special case of pattern recognition. Let $A(\delta) = \{x : \delta(x) = 1\}$. Since f_1 is a density that integrates to 1, we have:

$$R(\delta) = (1 - \pi) \int_{A(\delta)} f_0(x) dx + \pi(1 - \int_{A(\delta)} f_1(x) dx)$$

Which can be written as,

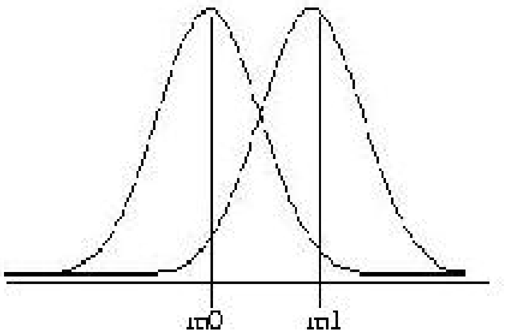
$$R(\delta) = \pi + \int [(1 - \pi) f_0(x) - \pi f_1(x)] 1_{A(\delta)}(x) dx$$

Notice that the rule claimed to be Bayes, assigns 1 when $\eta(x) > 1/2$ and this is equivalent to require the expression in square brackets above to be negative.

Now denote by $Q(x, \delta)$ the function that is being integrated above and show that for all $x \in R^d$, $Q(x, \delta^*) \leq Q(x, \delta)$ (just consider each of the four cases of x in or outside $A(\delta)$ and $A(\delta^*)$ separately). It then follows, by the last equation above and the monotonicity of integrals, that δ^* has the smallest possible probability of error and so it is the Bayes rule as claimed.

An Example

To make things concrete consider the simple case of deciding wether an observed real number x was generated by a gaussian with mean m_0 and variance 1 or a gaussian with mean m_1 and variance 1, and the two choices are considered a priori equally likely (see figure).



We want to compute the Bayes rule when the distribution of (X, Y) is characterized by $\pi = 1/2$ and $f_y = N(m_y, 1)$. For this case,

$$A(\delta^*) = \{x : f_0(x) < f_1(x)\} = \{x : \frac{-(x-m_0)^2}{2} < \frac{-(x-m_1)^2}{2}\}$$

and simplifies to,

$$A(\delta^*) = \{x : |x - m_1| < |x - m_0|\}$$

This shows that the best rule is to choose the distribution with mean closest to the observation x.

If we assume that $m_0 < m_1$ and let $m = (m_0 + m_1)/2$ be the middle point, then the Bayes risk R^* is,

$$R^* = \frac{1}{2}\int_m^\infty \varphi(x - m_0)dx + \frac{1}{2}\int_{-\infty}^m \varphi(x - m_1)dx$$

By letting $z = x - m_0$ in the first integral, $z = x - m_1$ in the second and defining $2a = (m_1 - m_0)$ we obtain a much simpler form,

$$R^* = \int_a^\infty \varphi(z)dz$$

depending only on the separation between the two means. Here are some values,

$\lvert m_1 - m_0 \rvert$	R^*
0.1	0.480
0.5	0.401
1.0	0.309
2.0	0.159
4.0	0.023
6.0	0.001

Learning Patterns from a Teacher

If we knew the distribution of (X, Y) we could compute δ^*, the Bayes rule, as shown above. Unfortunately, we rearly know the distribution of (X, Y). This creates two problems. First, we cannot directly use δ^* since it depends on the parameters of the unknown distribution of (X, Y). We need to know either the sign of $(\eta(x) - 1/2)$ or the sign of $[(1 - \pi)f_0(x) - \pi f_1(x)]$. Second, we can not even evaluate the risk $R(\delta)$ of any classifier δ for that also depends on the unknown distribution of (X, Y).

But if we have available n independent observations $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ of the vector (X, Y) then we could use them to estimate the missing distribution of (X, Y) with the empirical distribution based on the observations. This could be interpreted as past data that was labeled by a reliable teacher.

Thus, we could approximate the true probability $P\{(X, Y) \in C\}$ with the observed frequency given by the empirical measure ν_n,

$$\nu_n(C) = \frac{1}{n} \sum_{i=i}^{n} 1[(X_i, Y_i) \in C]$$

The true risk $R(\delta)$, for a given classifier δ, is estimated from the observed data with the empirical risk,

$$\hat{R}_n(\delta) = \frac{1}{n} \sum_{i=1}^{n} 1[\delta(X_i) \neq Y_i]$$

There is then the possibility of evaluating the quality of a δ with its observed frequency of errors $\hat{R}_n$ on the available data $D_n = (X_1, Y_1, \ldots, X_n, Y_n)$. But we should always keep in mind that this performance is different for different data sequences. A classifier that is built from given data D_n has as its true risk,

$$R_n = P\{\delta_n(X; D_n) \neq Y | D_n\}$$

Consistency

It is desirable to be able to quantify not just how good a classifier is on a fix data sequence D_n but to also know how a given classification method or *rule* behaves on most data sequences. In particular the notion of a *consistent* sequence $\{\delta_n\}$ of classifiers is simply the requirement that,

$$ER_n = P\{\delta_n(X; D_n) \neq Y\} \to R^* \quad \text{as } n \to \infty.$$

A classification rule may be consistent for some distributions of (X, Y) but not for others. If however it is consistent for all of them we say that it is *universally consistent*. Are there any known *universally consistent* rules? Yes! for example the k-nearest neighbor rule (*knn* for short) is one of them. The *knn* is simply the rule that assigns to x the most common observed label among its *knn*'s (in any metric topologically equivalent to the euclidean). Universal consistency for the *knn* is achieved only when $k = k(n)$ is such that $k \to \infty$ and $k/n \to 0$ as $n \to \infty$. These are the good news. The bad news: the convergence can be arbitrarily slow! Consistent rules with a given rate of convergence for all distributions of (X, Y) do not exist. This is interesting. On the one hand we know that given enough independent examples it is possible to get arbitrarily close to the performance of the Bayes rule. On the other hand it may take arbitrarily long for that to happen.

There are two ways out of this impasse. Use prior information to constraint the possible probability distributions for (X, Y) or change the target R^* by considering only a subclass $\mathcal{C}$ of classifiers. We take the second route first and show that there are classes $\mathcal{C}$ that provide universal rates of convergence to the best $\delta^*_{\mathcal{C}}$ for that class, i.e.,

$$R(\delta^*_{\mathcal{C}}) \leq R(\delta) \quad \text{for all} \quad \delta \in \mathcal{C}$$

Empirical Risk Minimization

Let us assume that we have data D_n as above from where we can estimate the true risk of a classifier δ by using the observed frequency of errors on D_n, i.e., by its empirical risk. Let δ^*_n be the classifier that minimizes the empirical risk over a given class $\mathcal{C}$ of rules, i.e.,

$$\hat{R}_n(\delta^*_n) \leq \hat{R}_n(\delta) \quad \text{for all } \delta \in \mathcal{C}$$

We would like to know what kinds of classes $\mathcal{C}$ allow universal empirical learning, in the sense that for all distributions of (X, Y),

$$R(\delta^*_n) \to R(\delta^*_{\mathcal{C}}) \quad \text{as } n \to \infty$$

It is intuitively clear that what's needed is some kind of constraint on the "size" (a better term would be capacity) of $\mathcal{C}$. For example if $\mathcal{C}$ contains all possible classification functions, then no matter what the sequence D_n is, we can always find a function in this class that fits the data perfectly. However we also feel intuitively that such a rule will over fit the observed data and it will show poor performance on sequences other than the observed D_n, i.e. we expect such rule to have poor generalization power. This intuitive reasoning will be rigorously confirmed by means of the celebrated Vapnik-Chervonenkis theory.

We first notice the following simple result:

Lemma 1

$$R(\delta_n^*) - \inf_{\delta \in \mathcal{C}} R(\delta) \leq 2 \sup_{\delta \in \mathcal{C}} |\hat{R}_n(\delta) - R(\delta)|$$

and

$$|\hat{R}(\delta_n^*) - R(\delta_n^*)| \leq \sup_{\delta \in \mathcal{C}} |\hat{R}_n(\delta) - R(\delta)|$$

Proof:

$$R(\delta_n^*) - \inf_{\delta \in \mathcal{C}} R(\delta) \;=\; R(\delta_n^*) - \hat{R}(\delta_n^*) + \hat{R}(\delta_n^*) - \inf_{\delta \in \mathcal{C}} R(\delta)$$

$$\leq \sup_{\delta \in \mathcal{C}} |R(\delta) - \hat{R}_n(\delta)| + \sup_{\delta \in \mathcal{C}} |\hat{R}_n(\delta) - R(\delta)|$$

The second part is trivial.

The above lemma shows that in order to achieve universal consistency on a given class, it is sufficient for the supremum appearing on the rhs of the inequalities to go to zero with probability 1, i.e., all we need to show is strong uniform convergence, over the given class, of empirical error to the true error.