ECE 901
Lecture 5: Plug-in Rules and the Histogram Classifier

R. Nowak

5/17/2009

We return to the topic of classification, and we assume an input (feature) space $\mathcal{X}$ and a binary output (label) space $\mathcal{Y} = \{0, 1\}$. Recall that the Bayes classifier (which minimizes the probability of misclassification) is defined by

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

Throughout this section, we will denote the conditional probability function by

$$\eta(x) \quad \equiv \quad P(Y = 1|X = x)$$

1 Plug-in Classifiers

One way to construct a classifier using the training data $\{X_i, Y_i\}_{i=1}^n$ is to estimate $\eta(x)$ and then plug-it into the form of the Bayes classifier. That is obtain an estimate,

$$\hat{\eta}_n(x) = \eta(x; \{X_i, Y_i\}_{i=1}^n)$$

and then form the "plug-in" classification rule

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

Remark: The function $\eta(x)$ is generally more complicated than the ultimate classification rule (binary-valued), as we can see

$$\begin{aligned} \eta &: \mathcal{X} \to [0, 1] \\ f &: \mathcal{X} \to \{0, 1\} \end{aligned}$$

Therefore, in this sense plug-in methods are solving a more complicated problem than necessary. Another way we can look at this is that we only need to estimate the $1/2$ *level set* of η, therefore it for some $x \in \mathcal{X}$ $\eta(x)$ is significantly far from $1/2$ we don't need to accurately estimate its value. Despite these draws however, plug-in methods can perform well, as demonstrated by the next result.

Theorem 1 (Plug-in Classifier) *Let $\tilde{\eta}$ be an approximation to η, and consider the plug-in rule*

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

Then,

$$R(f) - R^* \leq 2E[|\eta(x) - \tilde{\eta}(x)|]$$

where

$$\begin{aligned} R(f) &= P(f(X) \neq Y) \\ R^* &= R(f^*) = \inf_f R(f) \end{aligned}$$

Proof: In Lecture 2, we have shown that

$$R(f) - R^* = \int_{\mathcal{X}} |2\eta(x) - 1| \mathbf{1}\{f^*(x) \neq f(x)\} dP_X(x) \ ,$$

where f^* is the Bayes classifier (and $R^* = R(f^*)$).

Now note that

$$f(x) \neq f^*(x) \implies |\eta(x) - \tilde{\eta}(x)| \geq |\eta(x) - 1/2| \ .$$

This is easy to see by noticing that if $f(x) \neq f^*(x)$ then either $\eta(x) \geq 1/2$ and $\tilde{\eta}(x) < 1/2$ or $\eta(x) < 1/2$ and $\tilde{\eta}(x) \geq 1/2$. Using this fact we have

$$\begin{aligned} P(f(X) \neq Y) - R^* &= \int_{\mathcal{X}} 2|\eta(x) - 1/2| \, \mathbf{1}\{f^*(x) \neq f(x)\} dP_X(x) \\ &\leq \int_{\mathcal{X}} 2|\eta(x) - \tilde{\eta}(x)| \mathbf{1}\{f^*(x) \neq f(x)\} dP_X(x) \\ &\leq \int_{\mathcal{X}} 2|\eta(x) - \tilde{\eta}(x)| dP_X(x) \\ &= 2E[|\eta(X) - \tilde{\eta}(X)|] \ , \end{aligned}$$

where the second inequality is simply a result of the fact that $\mathbf{1}\{f^*(x) \neq f(x)\}$ is either 0 or 1.

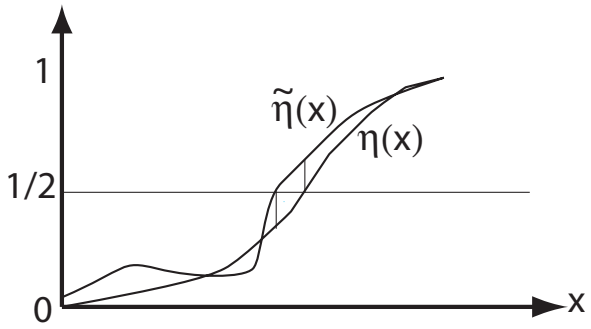

Figure 1: Pictorial illustration of $|\eta(x) - \tilde{\eta}(x)| \geq |\eta(x) - 1/2|$ when $f(x) \neq f^*(x)$. Note that the inequality $P(f(X) \neq Y) - R^* \leq \int_{\mathbf{R}^d} 2|\eta(x) - \tilde{\eta}(x)| \mathbf{1}\{f^*(x) \neq f(x)\} p_X(x) dx$ shows that the excess risk is at most twice the integral over the set where $f^*(x) \neq f(x)$. The difference $|\eta(x) - \tilde{\eta}(x)|$ may be arbitrarily large away from this set without effecting the error rate of the classifier. This illustrates the fact that estimating η well everywhere (i.e., regression) is unnecessary for the design of a good classifier (we only need to determine where η crosses the 1/2-level). In other words, "classification is easier than regression."

■

The theorem shows us that a good estimate of η can produce a good plug-in classification rule. By "good" estimate, we mean an estimator $\tilde{\eta}$ that is close to η in expected L_1-norm. Therefore if $E[|\hat{\eta}(X) - \eta(X)|] \to 0$ as $n \to \infty$ then $E[R(\hat{f}_n)] - R^* \to 0$ as $n \to \infty$, and so $\hat{f}_n$ is a consistent classifier (in the sense that when the amount n of training data increases the classifiers approaches the minimum classification error possible).

2 The Histogram Classifier

Let's assume that the (input) features are randomly distributed over the unit hypercube $\mathcal{X} = [0,1]^d$ (note that by scaling and shifting any set of bounded features we can satisfy this assumption), and assume that the (output) labels are binary, i.e., $\mathcal{Y} = \{0,1\}$. A histogram classifier is based on a partition the hypercube $[0,1]^d$ into M smaller cubes of equal size. Denote each one of these partition sets by Q_j, $j \in \{1,\ldots,m\}$.

Example 1 (Partition of hypercube in 2 dimensions) *Consider the unit square $[0,1]^2$ and partition it into M subsquares of equal area (assuming M is a squared integer). Let the subsquares be denoted by $\{Q_i\}$, $i = 1,\ldots,M$. See Figure 2 for an illustration.*

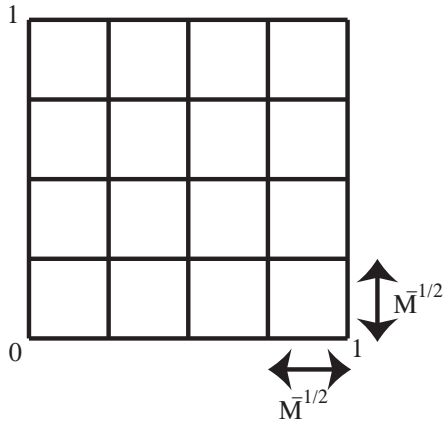

Figure 2: Example of hypercube $[0,1]^2$ in M equally sized partition

The idea now is to use this partition to construct of piecewise constant estimator for η. Define

$$\hat{\eta}_n(x) = \sum_{j=1}^{M} \hat{P}_j \mathbf{1}\{x \in Q_j\}$$

where

$$\hat{P}_j = \begin{cases} \frac{\sum_{i=1}^n \mathbf{1}\{X_i \in Q_j, Y_i = 1\}}{\sum_{i=1}^n \mathbf{1}\{X_i \in Q_j\}} & \text{if } \sum_{i=1}^n \mathbf{1}\{X_i \in Q_j\} \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

Like in the denoising example of Lecture 4, we expect that the bias of $\hat{\eta}_n$ will decrease as M increases, but the variance will increase as M increases. As before we will make M a function of n.

Let $\hat{f}_n(x) = \mathbf{1}\{\hat{\eta}_n(x) \geq 1/2\}$. We have the following result.

Theorem 2 (Consistency of Histogram Classifiers) *If $M_n \equiv M \to \infty$ and $\frac{n}{M_n} \to \infty$ as $n \to \infty$, then $E[|\hat{\eta}_n(X) - \eta(X)|] \to 0$ as $n \to \infty$, for every distribution P_{XY} with marginal density $p_X(x) \geq c$, for some constant $c > 0$.* [1].

As a consequence the plug-in classifier $\hat{f}_n$ is consistent, and so $E[R(\hat{f}_n)] - R^* \to 0$ as $n \to \infty$.

What the theorem tells us is that we need the number of partition cells to tend to infinity (to insure that the bias tends to zero), but they can't grow faster than the number of samples (*i.e.,* we want the number of samples per box tending to infinity to drive the variance to zero).

[1] Actually, the result holds for every distribution P_{XY}. For the more general theorem, refer to Theorem 6.1 in *A probabilistic Theory of Pattern Recognition* by Luc Devroye, László Györfi and Gábor Lugosi.

Proof: Let

$$P_j \equiv \frac{\int_{Q_j} \eta(x) p_X(x) dx}{\int_{Q_j} p_X(x) dx} \ .$$

This is the theoretical analog of $\hat{P}_j$. Define

$$\bar{\eta}(x) = \sum_{j=1}^{M} P_j \, \mathbf{1}\{x \in Q_j\} \ . \tag{1}$$

The function $\bar{\eta}$ is the theoretical analog of $\hat{\eta}$ (i.e., the function obtained by averaging η over the partition cells). By the triangle inequality,

$$E\left[|\hat{\eta}_n(X) - \eta(X)|\right] \ \leq \ \underbrace{E[|\hat{\eta}_n(X) - \bar{\eta}(X)|]}_{EstimationError} \ + \ \underbrace{E[|\bar{\eta}(X) - \eta(X)|]}_{ApproximationError}$$

Let's first bound the estimation error. Let $x \in [0,1]^d$ and let $j(x)$ denote the histogram bin in which x falls in (i.e. $x \in Q_{j(x)}$). To ease the notational burden consider a fixed (but arbitrary) point $x \in \mathcal{X}$ and drop the explicit dependence of j on x (in other words $j \equiv j(x)$). Define the random variables

$$N_j = \sum_{i=1}^{n} \mathbf{1}\{X_i \in Q_j\} \ ,$$

and

$$B_j = \sum_{i=1}^{n} Y_i \mathbf{1}\{X_i \in Q_j\} \ .$$

Note that

$$\hat{\eta}_n(x) = \begin{cases} \frac{B_j}{N_j} & \text{if } N_j \neq 0 \\ 0 & \text{otherwise} \end{cases} \ .$$

In other words N_j is the number of training examples that "landed" on cell Q_j and B_j is the number of samples in cell Q_j labelled 1. Note that Now $\hat{\eta}_n(x)$ is a fairly complicated random variable, but if we look separately at the distributions of N_j and B_j we can get some more information. First notice that $N_j \sim \text{Bin}(n, \int_{Q_j} dP_X(x)$, that is N_j is a *binomial* random variable. Notice also that if we consider the conditional distribution of B_j given N_j we have

$$B_j \,|\, N_j = k \ \sim \ \text{Binomial}\left(k, P(Y=1|X \in Q_j)\right) \ ,$$

where

$$\begin{aligned} P(Y=1|X \in Q_j) &= \frac{P(Y=1, X \in Q_j)}{P(X \in Q_j)} \\ &= \frac{\int_{Q_j} P(Y=1|X=x) dP_X(x)}{\int_{Q_j} dP_X(x)} \\ &= P_j \ . \end{aligned}$$

Now recall the definition of $\bar{\eta}$ (1). Fix $x \in \mathcal{X}$. Then

$$E[|\hat{\eta}(x) - \bar{\eta}(x)|] = E[|\hat{P}_{j(x)} - P_{j(x)}|] = E[|\hat{P}_j - P_j|] \ .$$

We can now use our previous observations and inspect the conditional expectation

$$E\left[|\hat{P}_j - P_j|\Big| \ N_j = k\right] \leq \begin{cases} E\left[|\hat{P}_j - P_j|\Big| \ N_j = k\right] & \text{if } k > 0 \\ \\ 1 & \text{if } k = 0 \quad (\text{since } 0 \leq \hat{P}_j, P_j \leq 1) \end{cases} \ .$$

For $k > 0$

$$\begin{aligned}
E\left[|\hat{P}_j - P_j|\Big|\ N_j = k\right] &= E\left[|\frac{B_j}{N_j} - P_j|\Big|\ N_j = k\right] \\
&= \frac{1}{k}E\left[|B_j - kP_j||\ N_j = k\right] \\
&= \frac{1}{k}E\left[|B_j - E[B_j|N_j = k]||\ N_j = k\right] \\
&\leq \frac{1}{k}\left(\underbrace{E\left[(B_j - E[B_j|N_j = k])^2\Big|\ N_j = k\right]}_{\text{conditional variance of } B_j}\right)^{1/2} \\
&= \frac{1}{k}\left(kP_j(1-P_j)\right)^{1/2} \\
&= \sqrt{\frac{P_j(1-P_j)}{k}} \\
&\leq \frac{1}{2\sqrt{k}}\ ,
\end{aligned}$$

where the first inequality is a consequence of Jensen's inequality, namely $E[|Z|] \leq (E[|Z|^2])^{\frac{1}{2}}$, and the second inequality follows by noting the worst case corresponds to $P_j = 1/2$.

In other words

$$\begin{aligned}
E\left[|\hat{\eta}(x) - \bar{\eta}(x)||\ N_{j(x)}\right] &\leq E\left[\left|\hat{P}_{j(x)} - P_{j(x)}\right|\Big|\ N_{j(x)}\right] \\
&\leq \begin{cases} \frac{1}{2\sqrt{N_{j(x)}}} & \text{if } N_{j(x)} > 0 \\ 1 & \text{if } N_{j(x)} = 0 \end{cases} \\
&= \frac{1}{2\sqrt{N_{j(x)}}}\mathbf{1}\{N_{j(x)} > 0\} + \mathbf{1}\{N_{j(x)} = 0\}\ .
\end{aligned}$$

As a consequence

$$\begin{aligned}
E\left[|\hat{\eta}(x) - \bar{\eta}(x)|\right] &= E\left[E\left[|\hat{\eta}(x) - \bar{\eta}(x)||\ N_{j(x)}\right]\right] \\
&= E\left[\frac{1}{2\sqrt{N_j}}\mathbf{1}\{N_j > 0\}\right] + P(N_j = 0) \\
&\leq \frac{1}{2}P(0 < N_j \leq \ell)\frac{1}{2\sqrt{\ell}}P(N_j > \ell) + P(N_j = 0) \\
&\leq \frac{1}{2}P(N_j \leq \ell) + \frac{1}{2\sqrt{\ell}}P(N_j > \ell) + P(N_j = 0)\ .
\end{aligned}$$

Where ℓ is an arbitrary integer, that we will choose carefully to optimize the bound. Now a key fact is that for any $\ell > 0$, $P(N_j \leq \ell) \to 0$ as $n \to \infty$. This follows from the assumption that the marginal density $p_X(x) \geq c$, for some constant $c > 0$, and $\frac{n}{M_n} \to \infty$ as $n \to \infty$. A bound on $P(N_j \leq \ell)$ can be easily computed using Chernoff's inequality (that we will see in a coming lecture), but for now just notice that $E[N_j] = nP(X \in Q_j) \geq nc\frac{1}{M}$, and so this expectation is going to infinity as n increases. Since N_j is a "well-behave" random variable this implies the result.

We are almost there. For any $\epsilon > 0$ there exists a $\ell > 0$ such that $\frac{1}{2\sqrt{\ell}} < \epsilon$ and $P(N_j \leq \ell) < \epsilon$ for n sufficiently large. Therefore, for n sufficiently large and every $x \in [0,1]^d$,

$$E\left[|\hat{\eta}_n(x) - \bar{\eta}(x)|\right] < 3\epsilon$$

where the expectation is with respect to the distribution of the sample $\{X_i, Y_i\}_{i=1}^n$. Thus,

$$E\left[|\hat{\eta}_n(X) - \bar{\eta}(X)|\right] \;<\; 3\epsilon$$

where the expectation is now with respect to the distribution of the sample and the marginal distribution of X.

The next and final step involves controling the approximation error $E\left[|\bar{\eta}_n(X) - \eta(X)|\right]$, where the expectation is over X alone. The function η may not itself be continuous, but there is another function η_ϵ that is uniformly continuous and such that $E\left[|\eta_\epsilon(X) - \eta(X)|\right] < \epsilon$ (This is a result from real-analysis, and we will not go into detail about it here. In essence what it means is that you can find a Lipschitz continuous function that is a good approximation to η). Recall that uniformly continuous functions can be well approximated by piecewise constant functions (as we have seen in Lecture 4).

By the triangle inequality,

$$E\left[|\bar{\eta} - \eta|\right] \leq \underbrace{E\left[|\bar{\eta} - \bar{\eta}_\epsilon|\right]}_{\leq \epsilon} + E\left[|\bar{\eta}_\epsilon - \eta_\epsilon|\right] + \underbrace{E\left[|\eta_\epsilon - \eta|\right]}_{\leq \epsilon \text{ by design}}$$

where $\bar{\eta}_\epsilon(x) = \sum_{j=1}^M \left[\frac{\int_{Q_j} \eta_\epsilon(x')dP_X(x')}{\int_{Q_j} dP_X(x')}\right] \mathbf{1}\{x \in Q_j\}$. Now

$$E\left[|\bar{\eta}(X) - \bar{\eta}_\epsilon(X)|\right] \leq \epsilon \;, \tag{2}$$

since $\bar{\eta}$ and $\bar{\eta}_\epsilon$ are "smoothed" versions of η and η_ϵ respectively (**Exercise:** formally show (2)). Finally since η_ϵ is uniformly continuous,

$$\begin{aligned}
E\left[|\bar{\eta}_\epsilon(X) - \eta_\epsilon(X)|\right] &= \sum_{j=1}^M \int_{Q_j} |\bar{\eta}_\epsilon(x) - \eta_\epsilon(x)| \mathbf{1}\{x \in Q_j\} p_X(x) dx \\
&\leq \sum_{j=1}^M \delta(M_n)\, P(x \in Q_j), \qquad \text{note that } \delta \text{ depends on } M \\
&= \delta(M_n), \qquad\qquad \text{since } \sum_{j=1}^M P(X \in Q_j) = 1
\end{aligned}$$

By taking M sufficiently large, δ can be made arbitrarily small. So for large M, $\delta \leq \epsilon$. And so we have shown that $E\left[|\bar{\eta} - \eta|\right] \leq 3\epsilon$ for large enough M, and given the conditions of the theorem

$$E\left[|\hat{\eta}(X) - \eta(X)|\right] \to 0$$

as $n \to \infty$, concluding the proof.

■