ECE901 Spring 2007 Statistical Learning Theory **Instructor: R. Nowak**

Lecture 5: Plug-in Classification Rules and Histogram Classifiers

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. 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: Consider any $x \in \mathbf{R}^d$. In proving the optimality of the Bayes classifier f^* in Lecture 2, we showed that

$$P\left(f(x) \neq Y|X = x\right) - P\left(f^*(x) \neq Y|X = x\right) \;=\; (2\eta(x) - 1)\left[\mathbf{1}_{\{f^*(x)=1\}} - \mathbf{1}_{\{f(x)=1\}}\right],$$

which is equivalent to

$$P\left(f(x) \neq Y | X = x\right) - P\left(f^*(x) \neq Y | X = x\right) \;=\; |2\eta(x) - 1| \; \mathbf{1}_{\{f^*(x) \neq f(x)\}},$$

since $f^*(x) = 1$ whenever $2\eta(x) - 1 > 0$. Thus,

$$\begin{aligned} P(f(X) \neq Y) - R^* &= \int_{\mathbf{R}^d} 2|\eta(x) - 1/2| \, \mathbf{1}_{\{f^*(x) \neq f(x)\}} p_X(x) dx \\ &\quad \text{where } p_X(x) \text{ is the marginal density of } X \\ &\leq \int_{\mathbf{R}^d} 2|\eta(x) - \tilde{\eta}(x)| \mathbf{1}_{\{f^*(x) \neq f(x)\}} p_X(x) dx \\ &\leq \int_{\mathbf{R}^d} 2|\eta(x) - \tilde{\eta}(x)| p_X(x) dx \\ &= 2E[|\eta(X) - \tilde{\eta}(X)|] \end{aligned}$$

where the first inequality follows from the fact

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

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

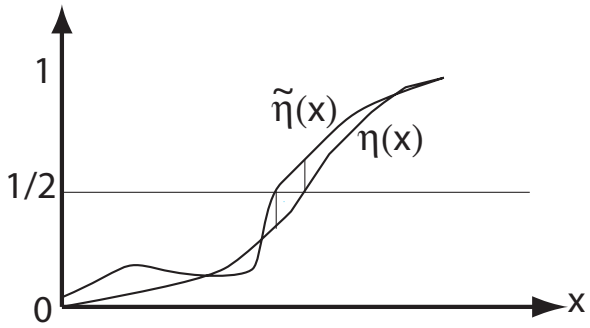

Figure 1: Pictorial illustration of $|\eta(x) - \tilde{\eta}(x)| \geq |\eta(x) - 1/2|$ when $f(x) \neq f^*(x)$. Note that the inequality $P(f(X) \neq Y) - R^* \leq \int_{\mathbf{R}^d} 2|\eta(x) - \tilde{\eta}(x)| \mathbf{1}_{\{f^*(x) \neq f(x)\}} p_X(x) dx$ shows that the excess risk is at most twice 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.

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.

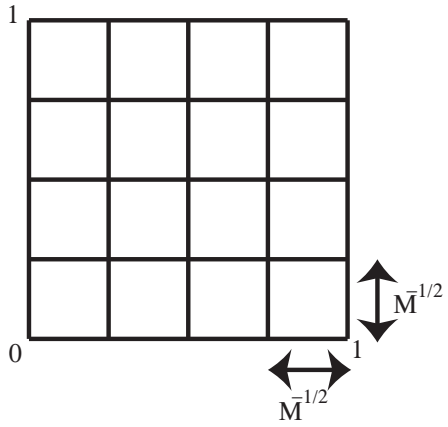

Figure 2: Example of hypercube $[0,1]^2$ in M equally sized partition

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 $\{Q_i\},\ i = 1,\ldots,M$.*

Define the following piecewise-constant estimator of $\eta(x)$:

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

where

$$\hat{P}_j = \frac{\sum_{i=1}^{n} \mathbf{1}_{\{X_i \in Q_j, Y_i = 1\}}}{\sum_{i=1}^{n} \mathbf{1}_{\{X_i \in Q_j\}}}.$$

Like our previous denoising examples, we expect that the bias of $\hat{\eta}_n$ will decrease as M increases, but the variance will increase as M increases.

Theorem 2 (Consistency of Histogram Classifiers) *If $M \to \infty$ and $\frac{n}{M} \to \infty$ as $n \to \infty$, then the histogram classifier risk converges to the Bayes risk for every distribution P_{XY} with marginal density $p_X(x) \geq c$, for some constant $c > 0$.*[1].

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).

Proof: Let $P_j \equiv \frac{\int_{Q_j} \eta(x) p_X(x) dx}{\int_{Q_j} p_X(x) dx}$ (the theoretical analog of $\hat{P}_j$) and define

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

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,

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

[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.

Let's first bound the estimation error. For any $x \in [0,1]^d$, let $Q(x)$ denote the histogram bin in which x falls in. Define the random variable

$$N(x) = \sum_{i=1}^{n} \mathbf{1}_{\{X_i \in Q(x)\}}$$

If $Q(x) = Q_j$, then this random variable is simply $n\hat{P}_j$. Note that

$$\hat{\eta}_n(x) = \frac{1}{N(x)} B(x)$$

where $B(x) == \sum_{i=1}^{n} \mathbf{1}_{\{X_i \in Q(x),\, Y_i = 1\}} = \sum_{i: X_i \in Q(x)} Y_i$. $B(x)$ is simply the number of samples in cell $Q(x)$ labelled 1. Now $\hat{\eta}_n(x)$ is a fairly complicated random variable, but the conditional distribution of $B(x)$ given $N(x)$ is relatively simple. Note that

$$B(x) \,|\, N(x) = k \;\sim\; \text{Binomial}\,(k, \bar{\eta}(x))$$

since $\bar{\eta}(x)$ is the probability of a sample in $Q(x)$ having the label 1 and we are conditioning on the event of observing k samples in $Q(x)$.

Now consider the conditional expectation

$$E\left[|\hat{\eta}_n(x) - \bar{\eta}(x)| \mid N(x) = k\right] \;\leq\; \begin{cases} E\left[\left|\frac{B(x)}{N(x)} - \bar{\eta}(x)\right| \mid N(x) = k\right], & k > 0 \\ 1, & k = 0 \quad (\text{since } 0 \leq \bar{\eta}(x) \leq 1) \end{cases}$$

Next note that

$$\begin{aligned} E\left[\left|\frac{B(x)}{N(x)} - \bar{\eta}(x)\right| \mid N(x) = k\right] &= E\left[\left|\frac{B(x)}{k} - \bar{\eta}(x)\right| \mid N(x) = k\right] \\ &= E\left[\frac{1}{k}|B(x) - \underbrace{k\bar{\eta}(x)}_{E[B(x)]}| \mid N(x) = k\right] \\ &\leq \frac{1}{k}\underbrace{\left(E\left[|B(x) - k\bar{\eta}(x)|^2 \mid N(x) = k\right]\right)}_{\text{conditional variance of } B(x)}{}^{\frac{1}{2}} \end{aligned}$$

by the Jensen's inequality, $E[|Z|] \leq (E[|Z|^2])^{\frac{1}{2}}$.

Therefore,

$$\begin{aligned} E\left[\left|\frac{B(x)}{N(x)} - \bar{\eta}(x)\right| \mid N(x) = k\right] &\leq \frac{1}{k}\left(k\bar{\eta}(x)(1 - \bar{\eta}(x))\right)^{\frac{1}{2}} \\ &= \sqrt{\frac{\bar{\eta}(x)(1 - \bar{\eta}(x))}{k}} \end{aligned}$$

and

$$E\left[|\hat{\eta}_n(x) - \bar{\eta}(x)| \mid N(x) = k\right] \;\leq\; \begin{cases} \sqrt{\frac{\bar{\eta}(x)(1-\bar{\eta}(x)}{k}}, & k > 0 \\ 1, & k = 0 \end{cases}$$

or in other words,

$$E\left[|\hat{\eta}_n(x) - \bar{\eta}(x)| \mid N(x) = k\right] \;\leq\; \sqrt{\frac{\bar{\eta}(x)(1 - \bar{\eta}(x)}{N(x)}}\mathbf{1}_{\{N(x)>0\}} \;+\; \mathbf{1}_{\{N(x)=0\}}$$

Now taking expectation with respect to $N(x)$

$$\begin{aligned} E_N\left[E[|\hat{\eta}_n(x) - \bar{\eta}(x)|N(x) = k]\right] &\leq E_N\left[\sqrt{\frac{\bar{\eta}(x)(1-\bar{\eta}(x)}{N(x)}}\mathbf{1}_{\{N(x)>0\}}\right] + P(N(x)=0) \\ &\leq E\left[\frac{1}{2\sqrt{N(x)}}\mathbf{1}_{\{N(x)>0\}}\right] + P(N(x)=0) \\ &\leq \frac{1}{2}P(N(x)\leq k) + \frac{1}{2\sqrt{k}}\underbrace{P(N(x)>k)}_{\leq 1} + P(N(x)=0) \end{aligned}$$

Now a key fact is that for any $k > 0$, $P(N \leq k) \to 0$ as $n \to \infty$. This follows from the assumption that the marginal density $p_X(x) \geq c$, for some constant $c > 0$, and $\frac{n}{M} \to \infty$ as $n \to \infty$. This result is easily verified by contradiction. If $P(N \leq k) \to q > 0$ as $n \to \infty$, then $P_X(x) > 0$ is contradicted. Thus, for any $\epsilon > 0$ there exists a $k > 0$ such that $\frac{1}{2\sqrt{k}} < \epsilon$ and $P(N \leq k) < \epsilon$ 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.

Next consider 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 $E\left[|\eta_\epsilon(X) - \eta(X)|\right] < \epsilon$. Recall that uniformly continuous functions can be well approximated by piecewise constant functions.

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 $\bar{\eta}_\epsilon(x) = \sum_{j=1}^m \left[\int_{Q_j} \eta_\epsilon(x')p_X(x')dx'\right]\mathbf{1}_{\{x\in Q_j\}}$.

$$\begin{aligned} E\left[|\bar{\eta}(X) - \bar{\eta}_\epsilon(X)|\right] &= \sum_{j=1}^m \left[\int_{Q_j} |\eta(x) - \eta_\epsilon(x)|p_X(x)dx\right]\mathbf{1}_{\{x\in Q_j\}} \\ &\leq \epsilon \end{aligned}$$

and since η_ϵ is uniformly continuous,

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

By taking M sufficiently large, δ can be made arbitrarily small. So for large M, $\delta \leq \epsilon$. Thus, we have shown

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

for sufficiently large M. Since $\epsilon > 0$ was arbitrary, we have shown that taking

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

satisfies

$$P(\hat{f}_n(X) \neq Y) - P(f^*(X) \neq Y) \;\leq\; 2E\left[|\hat{\eta}_n(X) - \eta(X)|\right] \to 0$$

if

$$\begin{array}{rcl} M & \to & \infty \\ \dfrac{n}{M} & \to & \infty \text{ as } n \to \infty \end{array}$$

Note: $P(\hat{f}_n(X) \neq Y) = E[\mathbf{1}_{\{\hat{f}(X) \neq Y\}}]$ is the expected risk of $\hat{f}$, with expectation over the distributions of (X, Y) and $\{X_i, Y_i\}_{i=1}^n$. ■