ECE901 Spring 2004 Statistical Regularization and Learning Theory **Lecture: 5**

The Histogram Classification Rule

Lecturer: Rob Nowak *Scribe: Yolanda Tsang*

1 Plug-in Classifiers

Recall that the Bayes classifier is defined by

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

where

$$\eta(x) \equiv P_{XY}(Y = 1|X = x)$$
$$x \in \mathbf{R}^d \text{ and } y = \{0, 1\}$$

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 plug-in classifier is generally more complicated, as we can see

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

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 $x \in \mathbf{R}^d$,

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

Recall from proof of optimality of f^* in Lecture 2. The above is equivalent to

$$\begin{aligned} & P\left(f(x) \neq Y | X=x\right) - P\left(f^*(x) \neq Y | X=x\right) \\ = \quad & |2\eta(x) - 1| \left(\mathbf{1}_{\{f^*(x) \neq f(x)\}}\right). \end{aligned}$$

Thus,

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

since

$$\begin{aligned} f(x) \neq f^*(x) \quad & \Longrightarrow \quad |\eta(x) - \tilde{\eta(x)}| \geq |\eta(x) - \frac{1}{2}| \\ f(x) = f^*(x) \quad & \Longrightarrow \quad P\left(f(x) \neq Y | X=x\right) = P\left(f^*(x) \neq Y | X=x\right) \end{aligned}$$

■

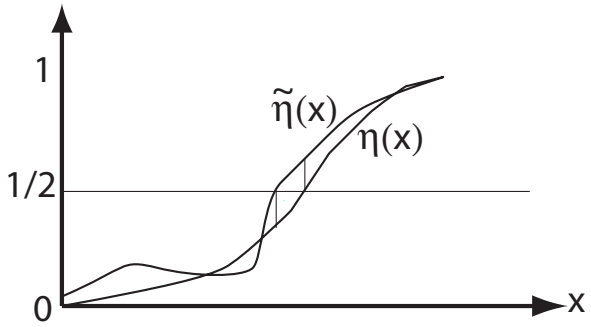

Figure 1: Pictorial illustration of $|\eta(x) - \tilde{\eta(x)}| \geq |\eta(x) - \frac{1}{2}|$ when $f(x) \neq f^*(x)$

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 features $\mathcal{X}$ are randomly distributed over the unit hypercube $[0,1)^d$. (Note that by scaling and shifting any set of bounded features we can satisfy this assumption).

Partition the hypercube $[0,1)^d$ into M smaller cubes of equal size.

Example 1 (Partition of hypercube in $2d$) *Consider the hypercube $[0,1]^2$ with M equally partitioned block. Let the cubes be denoted by $\{Q_i\},\ i = 1, \ldots, M$*

Estimate $\eta(x)$ by simply

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

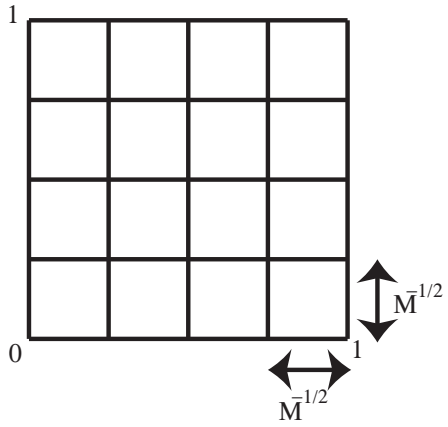

Figure 2: Example of hypercube $[0,1]^2$ in M equally sized partition

where

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

$\hat{\eta}_n$ is a piecewise constant estimator of η.

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 (Universally consistent) *If $M \to \infty$ and $\frac{n}{M} \to \infty$ as $n \to \infty$, then the histogram classifier is universally consistent*[1]

What the theorem says 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's consider what happens in one of the cubes $Q = Q_j$; for some j.

Define

$$\bar{\eta} = \sum_{j=1}^{M} \left[\int_{Q_j} \eta(x') P_X(x') dx' \right] \mathbf{1}_{\{x \in Q_j\}}$$

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

Let $P_j \equiv \int_{Q_j} \eta(x) P_X(x) dx$ (theoretical analog of $\hat{P}_j$)

By the triangle inequality,

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

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.

[1]Consistent for any distribution P_{XY} with $P_X(x) \geq c, c > 0 \forall x$. For a more general theorem, refer to Theorem 6.1 in *A probabilistic Theory of Pattern Recognition* by Luc Devroye, László Györfi and Gábor Lugosi.

Define the random variable

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

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

$$\hat{\eta}_n(x) = \frac{1}{N(x)} \sum_{i:X_i \in Q(x)} Y_i$$

Furthermore, if we condition on $N(x)$, then

$$N(x)\hat{\eta}_n(x) | N(x) = k \sim Binomial\,(k, \bar{\eta}(x))$$

Let $B(x) = N(x)\hat{\eta_n}(x)$. Then

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

Now,

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

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,

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

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

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

Since $\frac{n}{M}\to\infty$ as $n\to\infty$, it follows that for any $k>0$, $P(N\leq k)\to 0$ as $n\to\infty$. This follows 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 $x\in[0,1]^d$,

$$E_{D_n}\left[|\hat{\eta}_n(x)-\bar{\eta}(x)|\right]<2\epsilon$$

Now consider the approximation error $E\left[|\bar{\eta}_n(x)-\eta(x)|\right]$. 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 $\bar{\eta}_\epsilon(x)=\sum_{j=1}^m\left[\int_{Q_j}\eta_\epsilon(x')P_X(x')dx'\right]\mathbf{1}_{\{x\in Q_j\}}$.

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

and since η_ϵ is uniformly continuous,

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

Take $\delta=\epsilon$. Then

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

for sufficiently large M and

$$E\left[|\eta(x)-\hat{\eta}_n(x)|\right]\xrightarrow{P}0$$

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

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

if

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

■