Introduction to NonParametric Methods

1 The setup

Recall that there are 2 ways of describing the joint distribution of the pair $(\mathbf{X}, Y)$[1] Here, we first describe them, then comment on the forms of the Bayes Decision Rule and of the Bayes Risk.

1. Let $Y \sim Bern(\pi_1)^2$;
   Let $\mathbf{X}$ be conditionally distributed according to $f_Y(\mathbf{x})$ given Y.
2. Let $\mathbf{X} \sim f(\mathbf{X})$;
   Let Y be conditionally distributed as $Bern(\eta(\mathbf{x}))$ given Y.

Setup no 1 can be interpreted as follows:

- Spin a biased coin with probability of heads $= \pi_1$, if the toss yields heads, let Y be of class 1, otherwise of class 2;
- If $Y = 1$ sample $\mathbf{X}$ from $f_1(\mathbf{x})$, otherwise from $f_2(\mathbf{x})$.

Setup no 2 can be interpreted as follows:

- First, sample $\mathbf{X}$ from $f(\mathbf{x})$.
- Pick a biased coin, having probability of heads equal to $\eta(\mathbf{X})$, flip the coin, and if it yields heads, let $Y = 1$, otherwise let $Y = 0$.

Note that in setup no 2 we select the class label Y by selecting a biased coin from an infinite collection, using the observation $\mathbf{X}$.

Setup no 2 is probably less natural than setup no 1, however it has the remarkable advantage of relying directly on the posterior probability of class 1, $\eta(\mathbf{X})$, which is hidden in setup no 1, and must be obtained using Bayes Rule. The **form of the Bayes Decision Rule** for the fist setup is

[1] Recall that $\mathbf{X}$ is the observation and Y is the corresponding class label.

[2] We use the notation $Bern(p)$ to denote the Bernoulli distribution on a binary random variable Y, namely $\Pr\{Y = 1\} = p$, $\Pr\{Y = 0\} = 1 - p$.

- Decide $\hat{Y} = 1$ if $\pi_1 f_1(\mathbf{X}) > \pi_0 f_0(\mathbf{X})$[3]
- Decide $\hat{Y} = 0$ otherwise.

The form of the Bayes Decision Rule for the second setup is very appealing:

- Decide $\hat{Y} = 1$ if $\eta(\mathbf{x}) > 1/2$
- Decide $\hat{Y} = 0$ otherwise.

The reason why this is intuitively appealing is that it does not involve the distribution of $\mathbf{X}$. Of course, the two forms are equivalent, in the sense that they produce exactly the same decision regions.

The **form of the Bayes Risk** expressed in the first setup is[4]

$$R^* = \int_{\mathcal{X}} \min\left\{\pi_0 f_0(\mathbf{x}), \pi_1 f_1(\mathbf{x})\right\} \, d\mathbf{x},$$

while in the second setup is

$$R^* = \int_{\mathcal{X}} \min\left\{\eta(\mathbf{x}), 1 - \eta(\mathbf{x})\right\} f(\mathbf{x}) \, d\mathbf{x}, \tag{1}$$

2 Classifiers based on the second setup

A large family of classifiers estimate directly $\eta(\mathbf{x})$ rather than estimating the class conditional distributions and the priors, and applying the Bayes Rule to estimate $\eta(\mathbf{x})$.

Estimating $\eta(\mathbf{x})$ is a tricky business, and can have drawbacks: in particular, note that for classification we are really interested in estimating well $\eta(\mathbf{x})$ where its values are close to 1/2, while our estimates in regions where $\eta(\mathbf{x}) \sim 1$ or $\eta(\mathbf{x}) \sim 0$ need not be particularly accurate (to convince yourself of this, look at Equation 1).

At the same time, an algorithm that produces a good estimate $\hat{\eta}(\mathbf{x})$ of $\eta(\mathbf{x})$ yields good classification performance. More specifically, we can bound the error rate of a classifier $g(\mathbf{x})$ that produces the estimate $\hat{\eta}(\mathbf{x})$, and decides $\hat{Y} = g(\mathbf{X}) = 1$ if $\hat{\eta}(\mathbf{X}) > 1/2$ and $\hat{Y} = g(\mathbf{X}) = 0$ otherwise, in terms of the expected absolute difference between $\hat{\eta}(\mathbf{X})$ and $\eta(\mathbf{X})$.

Theorem *The risk R of the classifier $g(\mathbf{x})$ satisfies*

$$R \leq R^* + 2 \int_{\mathcal{X}} \left|\eta(\mathbf{x}) - \hat{\eta}(\mathbf{x})\right| f(\mathbf{x}) \, d\mathbf{x} \tag{2}$$

[3]Here π_0 is the prior probability of class 0 and $\pi_0 + \pi_1 = 1$.
[4]Recall that $\mathcal{X}$ is the feature space.

Proof Write the conditional probability of error of $g(\cdot)$ given $\mathbf{X} = \mathbf{x}$ as $\min\{\eta(\mathbf{x}), 1 - \eta(\mathbf{x})\} + \Delta\ell(\mathbf{x})$. There are 4 cases:

$\eta(\mathbf{x})$	$\hat{\eta}(\mathbf{x})$	$\Delta\ell(\mathbf{x})$	bound on $\Delta\ell(\mathbf{x})$	comment
$> 1/2$	$> 1/2$	0	$2\big(\eta(\mathbf{x}) - \hat{\eta}(\mathbf{x})\big)$	trivial
$> 1/2$	$\leq 1/2$	$2\eta(\mathbf{x}) - 1$	$2\big(\eta(\mathbf{x}) - \hat{\eta}(\mathbf{x})\big)$	see below
$\leq 1/2$	$> 1/2$	$1 - 2\eta(\mathbf{x})$	$2\big(\hat{\eta}(\mathbf{x}) - \eta(x)\big)$	see below
$\leq 1/2$	$\leq 1/2$	0	$2\big(\hat{\eta}(\mathbf{x}) - \eta(x)\big)$	trivial

Let's address first the case $\eta(\mathbf{x}) > 1/2$, $\hat{\eta}(\mathbf{x}) < 1/2$. Here we make the wrong decision and the conditional probability of error given $\mathbf{X} = \mathbf{x}$ is $\eta(\mathbf{x})$, while is it $1 - \eta(\mathbf{x})$ for the Bayes Decision Rule. The additional loss is therefore $\Delta\ell(\mathbf{x}) = 2\eta(\mathbf{x}) - 1 = 2(\eta(\mathbf{x}) - 1/2)$. Since, by assumption, $\hat{\eta}(\mathbf{x}) < 1/2$, clearly $\Delta\ell(\mathbf{x}) < 2\big(\eta(\mathbf{x}) - \hat{\eta}(\mathbf{x})\big)$. A similar argument applies to the other case.

Note that Equation 2 ensures that, if $\hat{\eta}(\mathbf{X})$ converges to $\eta(\mathbf{X})$ in the L_1 sense as the training set grows to infinity, then the risk of the rule generated by our classifier converges to the Bayes Risk. This is a powerful tool. In particular, if we can show convergence irrespective of the distribution of $(\mathbf{X}, Y)$, then we can prove **Universal Consistency** of the classification rule.

Def. *A classification rule $g_n(\cdot)$[5] is* **Universally Consistent** *if the risk R_n converges to R^* as the training set grows to ∞, for every distribution of the labeled samples $(\mathbf{X}, Y)$.*

---

3 The (cubic) Histogram Rule

The histogram rule estimates $\hat{\eta}(\mathbf{x})$ as follows

- Partition feature space into hypercubes $A_1, \ldots, A_m, \ldots$, of the same side h_n.
- Within each hypercube A_j: for each $\mathbf{x} \in A_j$ let $\hat{\eta}(\mathbf{x})$ be the ratio of the number of samples of class 1 in A_j to the number of all samples in A_j.

The classifier therefore assigns a label to each hypercube using majority vote; ties are broken by assigning the label 0.

Formally, let $A(\mathbf{X})$ be hypercube where $\mathbf{X}$ falls; then the cubic histogram rule is:

[5] Recall that a classification rule is a sequence of classifiers (i.e., hypothesis spaces and algorithms to learn hypotheses from the data) indexed by the training set size n.

- Decide 0 if $\sum_{i=1}^{n} Y_i 1_{\{X_i \in A(\mathbf{X})\}} \leq \sum_{i=1}^{n} (1 - Y_i) 1_{\{X_i \in A(\mathbf{X})\}}$
- Decide 1 otherwise.

3.1 Questions

- **What is good about this method?**
  - It is conceptually simple (visualize).
  - Gives rise to universally consistent rule.
  - We can easily prove theorems.
- **What is bad about this method?**
  - It suffers A LOT from the curse-of-dimensionality.
  - It usually is a poor performer in practice.

3.2 Universal Consistency of the Histogram Rule

A fundamental property of this seemingly simple rule is the following:

Theorem *If $h_n \to 0$ and $nh_n^d \to \infty$, the histogram rule is universally consistent.*

To prove this, we use a more general result: Consider generic rule $g_n^{'}(\mathbf{x})$ that divides the feature space into disjoint regions, or partitions, A_i and uses majority vote within each region. First, some notation

- *diam*$\{A\} = \sup_{\mathbf{x},\mathbf{z}} \|\mathbf{x} - \mathbf{z}\|$, corresponds to our intuitive notion of diameter, extended to nonclosed sets.
- $A(\mathbf{x})$ denote the region containing $\mathbf{x}$.
- $N(\mathbf{x}) = \sum_{i=1}^{n} 1_{\{\mathbf{X}_i \in A(\mathbf{x})\}}$ is the number of training samples falling in the partition that contains $\mathbf{x}$

Here is the main result:

Theorem *The rule $g_n^{'}(\mathbf{x})$ is universally consistent if*

- *diam*$\{A(\mathbf{X})\} \to 0$ *in probability.*
- $N(\mathbf{X}) \to \infty$ *in probability.*

This theorem states that if the probability that $\mathbf{X}$ falls in a region A, having vanishingly small diameter, and containing a number of training samples that grows to infinity, converges to 1 as the training set size goes to infinity, then $g_n^{'}(\mathbf{x})$ is universally consistent.

Proof We need some notation. Recall that $\eta(\mathbf{x}) = \Pr\{Y = 1 \mid \mathbf{X} = \mathbf{x}\}$. This quantity in general is not a constant within a region $A(\mathbf{x})$. The cubic histogram rule approximates $\eta(\mathbf{x}) = \Pr\{Y = 1 \mid \mathbf{X} = \mathbf{x}\}$ within $A(\mathbf{x})$ by a constant, which we denote by $\hat{\eta}_n(\mathbf{x})$, which is equal to

$$\hat{\eta}_n(\mathbf{x}) = N(\mathbf{x})^{-1} \sum_i Y_i 1_{\{\mathbf{X}_i \in A(\mathbf{x})\}} \tag{3}$$

if $N(\mathbf{x}) > 0$ and is equal to 0 otherwise. $\hat{\eta}_n(\mathbf{x})$ is really an estimator of the probability that a sample $\mathbf{X}$ has label 1 given that $\mathbf{X}$ falls within $A(\mathbf{x})$; this probability is denoted by $\tilde{\eta}(\mathbf{x})$ and is equal to

$$\tilde{\eta}(\mathbf{x}) = E_{\mathbf{X}}\left[\Pr\{Y = 1 \mid \mathbf{X} \in A(\mathbf{x})\}\right] = \frac{\int_{A(\mathbf{x})} \eta(\mathbf{x}^{'}) f(\mathbf{x}^{'})\, d\mathbf{x}^{'}}{\int_{A(\mathbf{x})} f(\mathbf{x}^{'})\, d\mathbf{x}^{'}}, \tag{4}$$

where $\mathbf{x}^{'}$ is the dummy variable for integration, and $f(\mathbf{x})$ is the density of $\mathbf{X}$.

We now bound $|\hat{\eta}_n(\mathbf{X}) - \eta(\mathbf{X})|$ in terms of $\tilde{\eta}(\mathbf{X})$:

$$E\left[|\hat{\eta}_n(\mathbf{X}) - \eta(\mathbf{X})|\right] \leq E\left[|\hat{\eta}_n(\mathbf{X}) - \tilde{\eta}(\mathbf{X})|\right] + E\left[|\tilde{\eta}(\mathbf{X}) - \eta(\mathbf{X})|\right] \tag{5}$$

where the inequality follows from the triangle inequality applied to the absolute value.

3.2.1 Bound on $E\left[|\hat{\eta}_n(\mathbf{X}) - \tilde{\eta}(\mathbf{X})|\right]$

Note that $N(\mathbf{x})\hat{\eta}_n(\mathbf{x})$ is $\sim \mathrm{Bin}(N(\mathbf{x}), \tilde{\eta}(\mathbf{x}))$ [6]. We will distinguish between the cases where $N(\mathbf{x}) = 0$ and $N(\mathbf{x}) > 0$.

Now, by definition, $\mathbf{X}$ falls in $A(\mathbf{X})$. Look at the indicators $1_{\{\mathbf{X}_i \in A(\mathbf{X})\}}$: these are equal to 1 if $\mathbf{X}_i \in$ same region as $\mathbf{X}$, and to 0 otherwise. Condition on $\mathbf{X}$ and on the indicators:

$$\begin{aligned} &E\left[|\hat{\eta}_n(\mathbf{X}) - \tilde{\eta}(\mathbf{X})| \;\middle|\; \mathbf{X}, 1_{\{\mathbf{X}_i \in A(\mathbf{X})\}}, i = 1, \ldots, n\right] \\ &\quad\leq E\left[\left|\frac{\mathrm{Bin}(N(\mathbf{X}), \tilde{\eta}(\mathbf{X}))}{N(\mathbf{X})} - \tilde{\eta}(\mathbf{X})\right| 1_{\{N(\mathbf{X})>0\}} \;\middle|\; \mathbf{X}, 1_{\{\mathbf{X}_1 \in A(\mathbf{X})\}}, i = 1, \ldots, n\right] \\ &\quad\quad + 1_{\{N(\mathbf{X})=0\}} \end{aligned} \tag{6}$$

[6]We use $\mathrm{Bin}(n, p)$ to denote the distribution of a binomial random variable with parameters n and p: this is the number of heads in n independent coin tosses of biased coins having probability of heads equal to p.

where the inequality follows from the facts that, when $N(\mathbf{X}) = 0$, $\hat{\eta}(\mathbf{X}) = 0$ and $E\left[\tilde{\eta}(\mathbf{X}) \mid \mathbf{X}, 1_{\{\mathbf{X}_i \in A(\mathbf{X})\}}, i = 1, \ldots, n\right] \leq 1$. Note: the thick $\mid$ is the conditioning sign, the thin ones indicate absolute value.

Now, recall the **Cauchy-Schwartz** inequality: if $E[U^2] < \infty$, $E[V^2] < \infty$,

$$E[|UV|] \leq \sqrt{E[U^2]\, E[V^2]}$$

In our case, $V = 1_{\{N(\mathbf{X})>0\}}$ (which can be moved in and out of the expected value), while U is the term inside the absolute value in Equation 6. Let's concentrate on V, and observe that the expected value of the first addend of V is equal to the second one;

$$E\left[\frac{\mathrm{Bin}\,(N(\mathbf{X}), \tilde{\eta}(\mathbf{X}))}{N(\mathbf{X})} \,\Bigg|\, \mathbf{X}, 1_{\{\mathbf{X}_i \in A(\mathbf{X})\}}, i = 1, \ldots, n\right] = \tilde{\eta}(\mathbf{X})$$

Hence $E[U^2]$ is the variance of the ratio, namely

$$\begin{aligned} E&\left[\left(\frac{\mathrm{Bin}\,(N(\mathbf{X}), \tilde{\eta}(\mathbf{X}))}{N(\mathbf{X})} - \tilde{\eta}(\mathbf{X})\right)^2 \,\Bigg|\, \mathbf{X}, 1_{\{\mathbf{X}_i \in A(\mathbf{X})\}}, i = 1, \ldots, n\right] \\ &= \mathrm{var}\left(\frac{\mathrm{Bin}\,(N(\mathbf{X}), \tilde{\eta}(\mathbf{X}))}{N(\mathbf{X})} \,\Bigg|\, \mathbf{X}, 1_{\{\mathbf{X}_i \in A(\mathbf{X})\}}, i = 1, \ldots, n\right) \end{aligned}$$

Recall that the variance of a Binomial(n, p) random variable is $np(1-p)$. Also, trivially, if $x = 0$ or 1, then $x^2 = x$, and therefore

$$E\left[\left(1_{\{N(\mathbf{X})>0\}}\right)^2 \,\Big|\, \mathbf{X}, 1_{\{\mathbf{X}_i \in A(\mathbf{X})\}}, i = 1, \ldots, n\right] = 1_{\{N(\mathbf{X})>0\}},$$

since, given the set of indicators, we know whether $N(\mathbf{X}) = 0$ or > 0, and the expected value of a constant is a constant. From these considerations and the Cauchy-Schwartz inequality, it follows that

$$\begin{aligned} E&\left[\left|\frac{B\,(N(\mathbf{X}), \tilde{\eta}(\mathbf{X}))}{N(\mathbf{X})} - \tilde{\eta}(\mathbf{X})\right| 1_{\{N(\mathbf{X})>0\}} \,\Bigg|\, \mathbf{X}, 1_{\{\mathbf{X}_i \in A(\mathbf{X})\}}, i = 1, \ldots, n\right] \\ &\leq E\left[\sqrt{\frac{\tilde{\eta}(\mathbf{X})\,[1 - \tilde{\eta}(\mathbf{X})]}{N(\mathbf{X})}} 1_{\{N(\mathbf{X})>0\}} \,\Bigg|\, \mathbf{X}, 1_{\{\mathbf{X}_i \in A(\mathbf{X})\}}, i = 1, \ldots, n\right] \quad (7) \end{aligned}$$

Note now the following:

- if $p \in [0, 1]$, then $\sqrt{p(1-p)} \leq 1/2$.
- Taking expectation with respect to $\mathbf{X}$ of $1_{\{N(\mathbf{X})=0\}}$, one obtains $\Pr\{N(\mathbf{X}) = 0\}$.

Hence, from the following considerations applied Equation 7, by taking the expectation with respect to $\mathbf{X}$ and $1_{\{X_i \in A(\mathbf{X})\}}$ of Equation 6, one obtains

$$\begin{aligned} E\left[\left|\hat{\eta}_n(\mathbf{X}) - \tilde{\eta}(\mathbf{X})\right|\right] &\leq E\left[\frac{1}{2\sqrt{N(\mathbf{X})}} 1_{\{N(\mathbf{X})>0\}}\right] + \Pr\{N(\mathbf{X}) = 0\} \\ &\leq \frac{1}{2}\Pr\{N(\mathbf{X}) \leq k\} + \frac{1}{2\sqrt{k}} + \Pr\{N(\mathbf{X}) = 0\} \quad (8) \end{aligned}$$

where Inequality 8 follows from

$$\begin{aligned} E\left[\frac{1}{\sqrt{N(\mathbf{X})}} 1_{\{N(\mathbf{X})>0\}}\right] &= \sum_{i=1}^{n} \frac{1}{\sqrt{i}} \Pr\{N(\mathbf{X}) = i\} \\ &= \sum_{i=1}^{k} \frac{1}{\sqrt{i}} \Pr\{N(\mathbf{X}) = i\} + \sum_{i=k+1}^{n} \frac{1}{\sqrt{i}} \Pr\{N(\mathbf{X}) = i\} \\ &\leq \sum_{i=1}^{k} \Pr\{N(\mathbf{X}) = i\} + \sum_{i=k+1}^{n} \frac{1}{\sqrt{i}} \Pr\{N(\mathbf{X}) = i\} \\ &= \Pr\{N(\mathbf{X}) \leq k\} + \sum_{i=k+1}^{n} \frac{1}{\sqrt{i}} \Pr\{N(\mathbf{X}) = i\} \\ &\leq \Pr\{N(\mathbf{X}) \leq k\} + \frac{1}{\sqrt{k}} \sum_{i=k+1}^{n} \Pr\{N(\mathbf{X}) = i\} \\ &= \Pr\{N(\mathbf{X}) \leq k\} + \frac{1}{\sqrt{k}} \Pr\{N(\mathbf{X}) > k\} \\ &\leq \Pr\{N(\mathbf{X}) \leq k\} + \frac{1}{\sqrt{k}} \end{aligned}$$

But, by assumption, for every $k > 0$ $\Pr\{N(\mathbf{X}) > k\} \to 1$, so the first and third terms of Equation 8 are arbitrarily small (say $> \epsilon/4$ for any chosen ϵ) for large enough n, and the second term can be made small by choosing k sufficiently large (again, say $> \epsilon/4$).

3.2.2 Bound on $E\left[\left|\tilde{\eta}(\mathbf{X}) - \eta(\mathbf{X})\right|\right]$

In the previous subsection, we had the remarkable advantage of dealing with piecewise constant functions ($\hat{\eta}(\mathbf{X})$ and $\tilde{\eta}(\mathbf{X})$). Here we are dealing with $\eta(\mathbf{X})$, which is a potentially very "unsmooth" function (it need not be continuous, for example, and it could be discontinuous everywhere.....).

A standard approach in these cases is to find a nicely behaved function that approximates $\eta(\mathbf{X})$ to a desired degree, apply the triangle inequality to the absolute value, and spend most of our effort dealing with it.

Claim: we can find a function $\eta_\epsilon(\mathbf{x})$ with the following properties:

- it is $[01]-$valued,
- vanishes off a set C,
- it is uniformly continuous on C
- it satisfies $E\left[\mid \eta(\mathbf{X}) - \eta_\epsilon(\mathbf{X}) \mid\right] \leq \epsilon/6$.

The interested reader is referred to Devroye, Györfi, and Lugosi's book for details on why we can actually find such function.

Define $\tilde{\eta}_\epsilon$ as its expectation (as above), more specifically, let

$$\tilde{\eta}_\epsilon(\mathbf{x}) = E\left[\eta_\epsilon(\mathbf{X}) \mid \mathbf{X} \in A(\mathbf{x})\right]$$

Again, we apply the triangle inequality; Since now we have four functions ($\eta(\mathbf{X})$, $\tilde{\eta}(\mathbf{X})$, $\eta_\epsilon(\mathbf{X})$, and $\tilde{\eta}_\epsilon(\mathbf{X})$), the triangle inequality yields 3 terms

$$\begin{aligned} E\left[|\tilde{\eta}(\mathbf{X}) - \eta(\mathbf{X})|\right] &\leq E\left[|\tilde{\eta}(\mathbf{X}) - \tilde{\eta}_\epsilon(\mathbf{X})|\right] + E\left[|\tilde{\eta}_\epsilon(\mathbf{X}) - \eta_\epsilon(\mathbf{X})|\right] \\ &\quad + E\left[|\eta_\epsilon(\mathbf{X}) - \eta(\mathbf{X})|\right] = I + II + III \end{aligned}$$

Now: $III < \epsilon/6$ by construction.

To bound II:

- recall what uniform continuity means: for every ϵ, $\delta(\epsilon)$ such that if $diam\{A(\mathbf{X})\} \leq \delta(\epsilon)$, then

  $$\sup_{\mathbf{x}_1, \mathbf{x}_2 \in A(\mathbf{X})} |\eta_\epsilon(\mathbf{x}_1) - \eta_\epsilon(\mathbf{x}_2)| \leq \epsilon/6,$$

  irrespective of where $A(\mathbf{X})$ is[7]. Since the range of $\eta_\epsilon(\mathbf{x})$ is limited by $\leq \epsilon/12$ (the 12 will come handy later) in $A(\mathbf{X})$, a fortiori its expected deviation from its mean within $A(\mathbf{X})$ is bounded by $\epsilon/6$, hence $II < \epsilon$.
- Also note that: $II \leq 1$, irrespective of $A(\mathbf{X})$

Hence:

$$\begin{aligned} II &\leq \frac{\epsilon}{12} \Pr\left\{diam\{A(\mathbf{X})\} \leq \delta(\epsilon)\right\} + 1 \cdot \Pr\left\{diam\{A(\mathbf{X})\} > \delta(\epsilon)\right\} \\ &\leq \frac{\epsilon}{12} + \Pr\left\{diam\{A(\mathbf{X})\} > \delta(\epsilon)\right\} \\ &\leq \frac{\epsilon}{6} \end{aligned}$$

[7] Provided that its diameter is $< \delta(\epsilon)$

for n large enough, since by assumption $diam\{A(\mathbf{X})\} \to 0$ in probability. Finally:

$$\begin{aligned} III &= E\left[\left|\eta_\epsilon(\mathbf{X}) - \eta(\mathbf{X})\right|\right] \\ &= \sum_i \Pr\{\mathbf{X} \in A_i\} E\left[\left|\eta_\epsilon(\mathbf{X}) - \eta(\mathbf{X})\right| \,\Big|\, \mathbf{X} \in A_i\right] \end{aligned}$$

We now rely on a fundamental inequality: **Theorem Jensen's Inequality** *If f is an integrable, convex function,*

$$E[f(X)] \geq f(E[X])$$

Note that the absolute value is a convex function. Note also that

$$\begin{aligned} E[|\tilde{\eta}(\mathbf{X}) - \tilde{\eta}_\epsilon(\mathbf{X})|] &= \sum_i \Pr\{X \in A_i\} \\ &\quad \cdot \left| E[\Pr\{Y = 1 \mid \mathbf{X} \in A(\mathbf{x})\}] - E[\eta_\epsilon(\mathbf{X}) \mid \mathbf{X} \in A(\mathbf{x})] \right| \end{aligned}$$

hence $I \leq III < \epsilon/6$ by Jensen's and the convexity of $|\cdot|$. Adding all the results we conclude that the LHS of Equation 6 is less than ϵ

For the histogram rule: the diameter of the cell is $\sqrt{d}h_n$: by assumption $\to 0$

Lemma For the histogram rule, the probability that $N(\mathbf{X}) < M$ goes to zero.
(No proof, this is somewhat complex. The basic idea of the proof is the following: put large ball around origin, divide the cells into cells intersecting the ball and cells not intersecting the ball. Bound probability that $N(\mathbf{X}) > M$ for all nonintersecting cells with 1, and therefore the contribution of all nonintersecting cells is less than the probability that $\mathbf{X}$ falls outside the ball; this probability can be made arbitrarily small by selecting a large enough ball.

For intersecting cells, count them (their number goes to ∞ as $1/h_n^d$). Divide them into cells with probability $< 2M/n$ and into cells with probability $\geq 2M/n$ (with respect to the law of $\mathbf{X}$).

Bound the sum over the first cells (with small probability) by the overall number of intersecting cells times their probability (this yields nh_n^d in the denominator). For the remaining ones, the expected number of counts is $n\mu(A)$, so they have a lot of points. The probability that the observed number of counts minus $n\mu(A)$ is $\leq M - n\mu(A)$ is bounded using Chebycheff inequality, (the variance is $n * [n\mu(A)][n(1 - \mu(A))]$))

---

4 Kernel Rules

A problem with histogram rule is the fact that all points within a cell have the same "voting power" irrespective of where $\mathbf{X}$ falls within the cell. To deal with this problem, we could use a moving window, rather than a fixed partition of the space: for example, we could use a sphere of radius r centered at $\mathbf{X}$.

This looks better, but still points near the edge of the window count as much as points near the center.

SOLUTION: Filter !

Let $K(\mathbf{x})$ be a "kernel function" (we will define what we actually mean later). At each $\mathbf{x}$ compute [8]

$$\hat{\eta}_n(\mathbf{x}) = \frac{\sum_{i=1}^{n} 1_{\{Y_i==1\}} K(\mathbf{X}_i - \mathbf{x})}{\sum_{i=1}^{n} K(\mathbf{X}_i - \mathbf{x})},$$

called the **Nadaraya-Watson kernel-weighted average**.

In what follows, we use the notation $K_h(\mathbf{x}) = h^{-1}K(\mathbf{x}/h)$. The parameter h controls the width (and height) of the kernel. Larger values of h correspond to a wider, more spread version of the kernel, while small values of h correspond to concentrated kernels

4.1 Selecting kernels

A "good" kernel is a function that gives rise to universally consistent rules.

Def. Regular Kernel A regular kernel is a kernel satisfying the following condition:

- it is nonnegative
- it is bounded
- it is integrable with respect to the Lebesgue measure
- it is uniformly continuous

Theorem Universal Consistency *If*

- *K is regular;*

[8] The denominator is really $\sum_{i=1}^{n} 1_{\{Y_i==1\}} K(\mathbf{X}_i - \mathbf{x}) + \sum_{i=1}^{n} 1_{\{Y_i==0\}} K(\mathbf{X}_i - \mathbf{x})$

- $h \to 0$
- $nh^d \to \infty$ *as* $n \to \infty$

then for EVERY $P(\mathbf{X}, Y)$, *for every* $\epsilon > 0$, $\exists n_\epsilon > 0$ *s.t.,* $\forall n > n_\epsilon$

$$Pr\{R_n - R^* > \epsilon\} \leq 2e^{-n\epsilon^2/32\rho}$$

where $\rho > 0$ *depends on* K *and* d *only* [9].

4.2 Examples of Kernels

We concentrate the attention to d-dimensional symmetric kernels derived from $1-$dimensional kernels: if $K(x)$ is a $1-$d kernel, $C_d K(\|\mathbf{x}\|)$ is a $d-$Dim, symmetric kernel, C_d: normalizing constant.

- Rect. Window: $K(x) = 1_{\{|x| \leq 1\}}$
- Triangular: $K(x) = 1_{\{|x| \leq 1\}}(1 - |x|)$
- Epanechnikov: $K(x) = \frac{3}{4} 1_{\{|x| \leq 1\}}(1 - x^2)$
- Tri-cube: $K(x) = 1_{\{|x| \leq 1\}}(1 - x^3)^3$ flatter on top
- Gaussian: $K(x) = \exp(-x^2/2)/\sqrt{2\pi}$: ∞ support
- Bell: $K(x) = 1_{\{|x| \leq 1\}} \exp(-1/(1-x)^2)$: infinitely differentiable + compact support.

4.2.1 Which one is better?

Hard to tell. However,

- Smoother k's produce smoother $\hat{\eta}$
- Compactly supported k's computationally cheaper
- In practice, it makes little difference

4.3 So what is the important parameter?

The choice of the parameter λ. How do we select λ? In class we saw an example demonstrating that it is really hard to gauge from visual inspection of the data what a good choice of λ is. So? We propose here two approaches

Test set method

[9] The Borel Cantelli Lemma assures that the probability that $R_n - R^* > \epsilon$ infinitely often is zero.

- Divide data in training set+test set.
- Select a "candidate range" for λ (a good range is in general related to the average distance between training points: you want a range so that for each λ in the range, for most values of $\mathbf{x}$, the kernel utilizes a reasonable number of training points. This can be estimated by centering the kernel at each of the training points, and counting how many other training points receive a weight larger than some threshold ϵ);
- Divide the range of "candidate" λ into k val's $\lambda_1 \ldots \lambda_k$;
- For each λ_i, compute error rate on test set;
- Select the best λ_i.

This algorithm is simple. It could be improved as follows:

Cross-Validation

- Randomly split data into M groups of same size.
- Divide range of "candidate" λ into k val's $\lambda_1 \ldots \lambda_k$.
- For $j = 1, \ldots, M$, let ith group be test set, and the rest of the data as training set
- Repeat the test set procedure using the training and test data described in the previous step; remember the error rates.
- Select λ_i with minimum Average error.

4.4 Weighted Regression

We briefly describe another interesting related method. The weighted regression works only in low-dimensional spaces, and is a method for improving the estimate of $\eta(\mathbf{x})$.

Recall the MSE procedure described in class. Note that it was equivalent to fitting a line to the data that best approximates the labels. Again, let vectors be row vectors.

Now, unlike the MSE procedure that finds $\mathbf{w}^*$ minimizing $\sum_i (Y_i - \mathbf{w}\mathbf{X}_i^T)^2$, at each $\mathbf{x}$ we compute $\mathbf{w}^*(\mathbf{x})$ minimizing

$$\sum_i K(\mathbf{X}_i - \mathbf{x}) \left(Y_i - \mathbf{w}(\mathbf{x})\mathbf{X}_i^T\right)^2$$

In matrix form: arrange the $\mathbf{X}_i$ in the rows of a matrix, add a first column with all 1's; call the matrix B Let $W(\mathbf{x})$ be the $N \times N$ diagonal matrix, with ith element equal to $K_\lambda(\mathbf{x}, \mathbf{X}_i)$. Then

$$\hat{\eta}(\mathbf{x}) = [1\ \mathbf{x}] \left(B^T W(\mathbf{x}_0) B\right)^{-1} B^T W(\mathbf{x}) \mathbf{Y} = \sum_{i=1}^{N} l_i(\mathbf{x}) Y_i$$