ECEN 649 Pattern Recognition

Nonparametric Classification Rules

Ulisses Braga-Neto

ECE Department

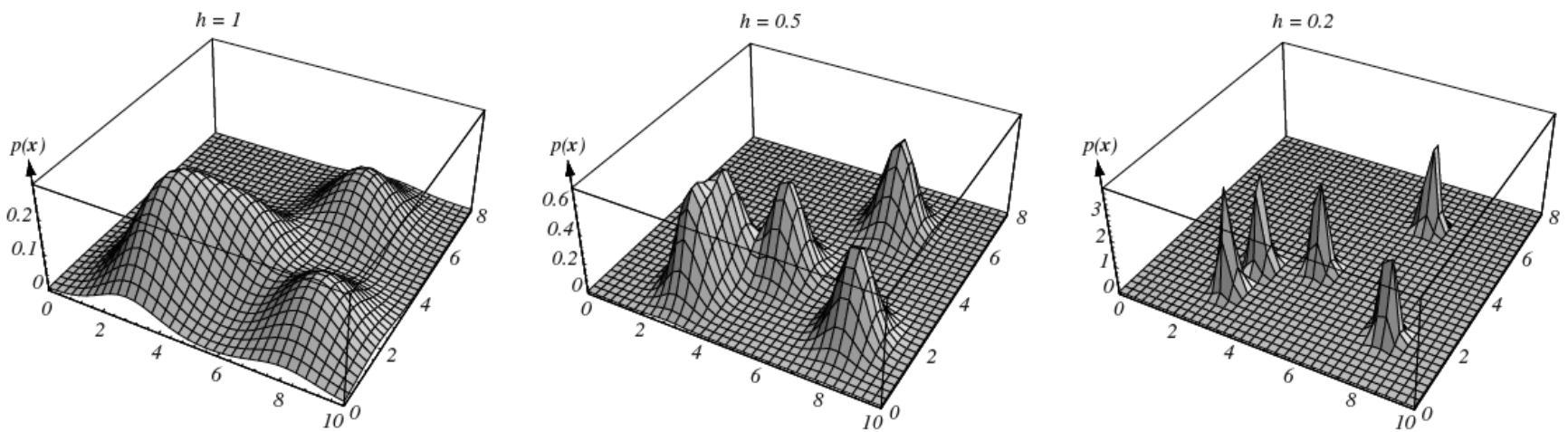
Texas A&M University

Basic Idea

- The basic idea of nonparametric classification is to estimate $\eta(x)$ (or the a-priori probabilities and class-conditional densities) directly, without making any distributional assumptions, and then plug into the expression for the Bayes classifier.
- Intuitively, this has a better chance of producing universally consistent classification rules (and it does).
- However, one should always keep in mind that the small-sample performance has little to do with consistency (some nonparametric rules do badly in small-sample settings).

Smoothing

- All nonparametric rules derive an estimate of $\eta(x)$ by smoothing the data. The drawback is that the amount of smoothing must be specified. The selection of this "smoothing parameter" may be difficult (this is the problem of model selection, which we will discuss later).
- Example: "Parzen windows"



Histogram Rules

- Histogram rules are based on *partitions* of the feature space. A partition is a mapping $A : R^p \rightarrow \mathcal{P}(R^p)$ such that

$$A(x) = A(y) \text{ or } A(x) \cap A(y) = \emptyset$$

and

$$R^d = \bigcup_{x \in R^p} A(x)$$

- The estimate $\eta_n(x)$ of $\eta(x)$ is derived as:

$$\eta_n(x) = \frac{1}{N(x)} \sum_{X_i \in A(x)} Y_i$$

where $N(x) = \#\{X_i | X_i \in A(x)\}$.

Histogram Rules - II

- The designed classifier is given by:

$$\psi_n(x) = \begin{cases} 1, & \eta_n(x) > \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

$$= \begin{cases} 1, & \sum_{X_i \in A(x)} I_{\{Y_i=1\}} > \sum_{X_i \in A(x)} I_{\{Y_i=0\}} \\ 0, & \text{otherwise} \end{cases}$$

- In other words, $\psi_n(x)$ is 1 if the zone $A(x)$ of the partition containing x contains more sample points with label 1 than label 0.
- This is therefore called a *majority vote* classifier.

Consistency of Histogram Rules

- Let $\{A_n | n \geq 1\}$ be a sequence of partitions indexed by the sample size n, and define $N_n(x)$ and ψ_n acoordingly, for each A_n. The condition for this classification rule to be consistent is given by the next theorem.

- (DGL Theorem 6.1) For the above rule, $E[\epsilon_n] \rightarrow \epsilon^*$ if

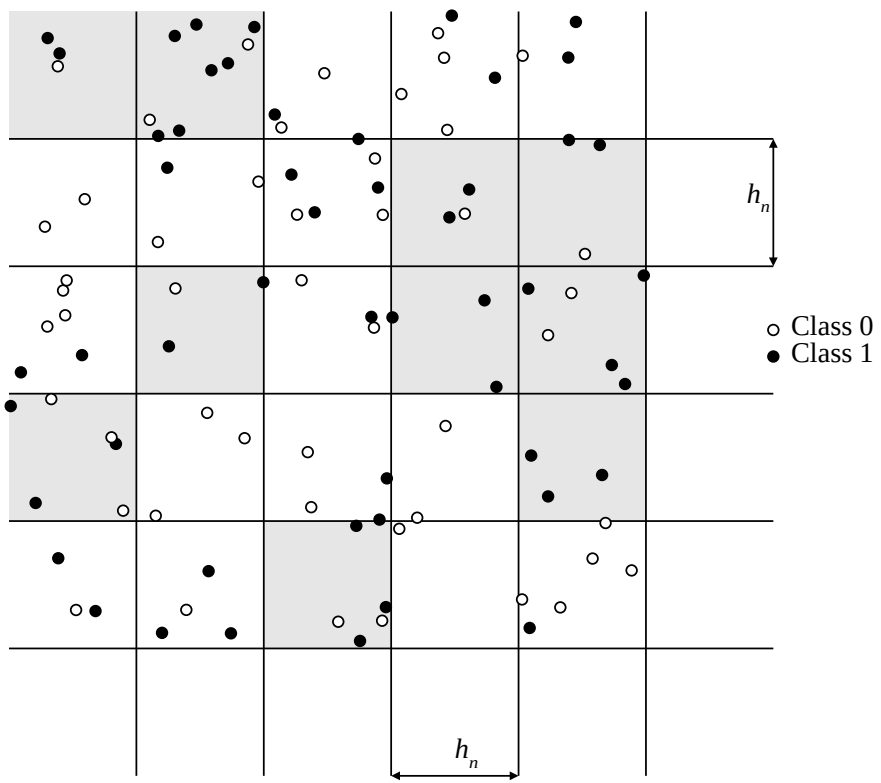
  (i) diam$[A_n(X)] = \sup_{x,y \in A_n(x)} ||x - y|| \rightarrow 0$ in probability

  (ii) $N_n(X) \rightarrow \infty$ in probability

  Idea of proof: show that $E[|\eta_n(X) - \eta(X)|] \rightarrow 0$.

Cubic Histogram Rule

- This is the special case where each zone of the partition A_n is a cube of side h_n, organized in a regular grid.
- Bi-dimensional example.



Universal Consistency

- (DGL Theorem 6.2) Let $V_n = h_n^d$ be the common volume of all cells. If $h_n \to 0$ (so $V_n \to 0$) but $nV_n \to \infty$ as $n \to \infty$, then $\epsilon_n \to \epsilon^*$ and the cubic histogram rule is universally consistent.

  Idea of proof: Show that conditions of DGL Theorem 6.1 hold for any distribution F_{XY}. Part (i) is easy: diam$[A_n(X)] = \sqrt{d}h_n$ (= cube diagonal) $\to 0$. Then show part (ii) by proving that for any M, $P(N_n(X) < M) \to 0$.

- DGL Theorem 9.4 shows universal strong consistency.

- Finite sample performance: how to pick the smoothing parameter h_n? This is a model selection problem (more on this later).

Nearest Neighbor Rules

- Given $x \in R^d$, let

$$(X_{(i)}(x), Y_{(i)}(x)) \;=\; i\text{-th nearest data point to x,}$$

for $i = 1, \ldots, n$.

- The estimate $\eta_n(x)$ of $\eta(x)$ is derived as:

$$\eta_n(x) \;=\; \frac{1}{K} \sum_{i=1}^{K} Y_{(i)}(x)$$

where K is a specified integer parameter.

Nearest Neighbor Rules - II

- The designed classifier is given by:

$$\psi_n(x) = \begin{cases} 1, & \eta_n(x) > \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

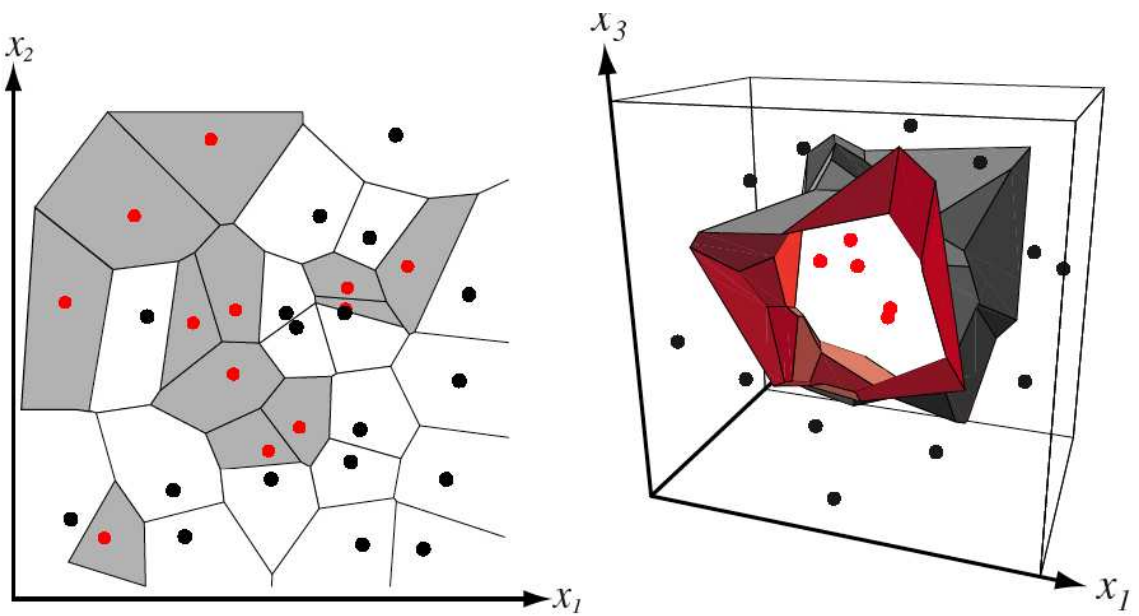
$$= \begin{cases} 1, & \sum_{i=1}^{K} I_{\{Y_i=1\}} > \sum_{i=1}^{K} I_{\{Y_i=0\}} \\ 0, & \text{otherwise} \end{cases}$$

- In other words, $\psi_n(x)$ is 1 if the among the K nearest neighbors of x there are more labels 1 than labels 0.

- This is called the K-*nearest neighbor* (KNN) classifier.

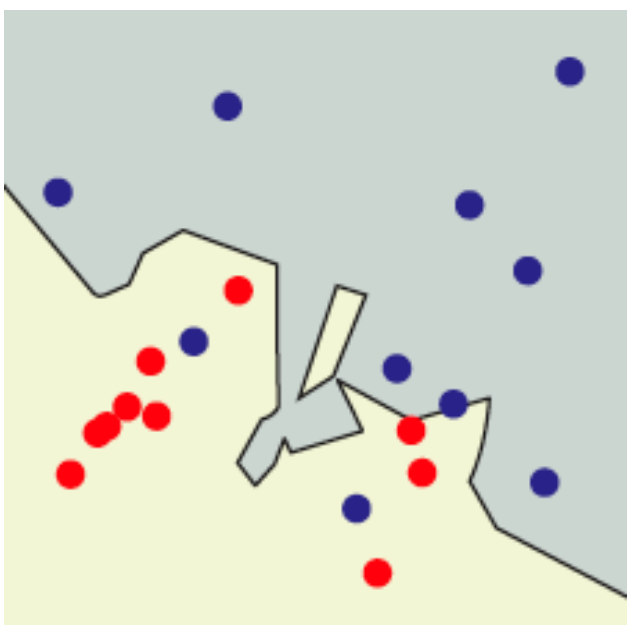
Nearest Neighbor Classifier (1NN)

- The case $K = 1$ is of considerable historical interest. In practice, it does not perfrom well with small-samples because it tends to overfit the data.
- Its decision surface is very complex (Voronoi diagram).
- Example: 2-D (left) and 3-D (right).



3-Nearest Neighbor Classifier (3NN)

- The case $K = 3$ shows in practice, especially with small samples, to be a good compromise between too little smoothing ($K = 1$) and too much smoothing ($K \geq 5$).
- Example.



Universal Consistency

- (DGL Theorem 6.4) If $K \to \infty$ while $K/n \to 0$ as $n \to \infty$, then for all distributions $\epsilon_n \to \epsilon^*$ and the KNN rule is universally consistent.
- This theorem is due to C. Stone (1977) and was the first proof of universal consistency for any classification rule.
- DGL Theorem 11.1 shows universal strong consistency assuming X has a density.
- Finite-sample performance: how to pick the smoothing parameter K? Again, this is a model selection problem (more on this later).

Fixed-K Asymptotic Performance

- In practice, K is kept fixed. Is the rule universally consistent then?
- The answer is no. We will see however that

$$\epsilon_{KNN} = \lim_{n \to \infty} E[\epsilon_n] \leq a_K \epsilon^*$$

where $a_K > 1$ is a constant that depends on K (note that this implies consistency if $\epsilon^* = 0$).

Cover-Hart Theorem

- One of the most famous theorems of Pattern Recognition was proved by Cover and Hart in 1967. It implies that the asymptotic performance of the nearest neighbor (1NN) rule is at worst twice the Bayes error.
- (DGL Theorem 5.1) For any distribution F_{XY}, one has

$$\epsilon_{NN} = E[2\eta(X)(1-\eta(X))]$$

- As can be shown easily (DGL p. 22), this implies

$$\epsilon_{NN} \leq 2\epsilon^*(1-\epsilon^*) \leq 2\epsilon^*$$

so that $a_1 = 2$.

Extension of Cover-Hart Theorem

- For general odd K, one has

$$\epsilon_{KNN} = E\left[\eta(X)P\left(B(K,\eta(X)) < \frac{K}{2}\,\middle|\,X\right) + (1-\eta(X))P\left(B(K,\eta(X)) > \frac{K}{2}\,\middle|\,X\right)\right]$$

where $B(n,p)$ denotes a binomial random variable with parameters n and p, so that:

$$P\left(B(K,\eta(X)) < \frac{K}{2}\,\middle|\,X\right) = \sum_{i=0}^{(K-1)/2}\binom{K}{i}\eta^i(X)(1-\eta(X))^{K-i}$$

with a similar expression for $P\left(B(K,\eta(X)) > \frac{K}{2}\,\middle|\,X\right)$.

Extension of Cover-Hart Theorem - II

- For example, for $K = 3$, one gets

$$\epsilon_{3NN} = E\left[\eta(X)\left(\binom{3}{0}(1-\eta(X))^3 + \binom{3}{1}\eta(X)(1-\eta(X))^2\right)\right.$$
$$\left. + (1-\eta(X))\left(\binom{3}{2}\eta(X)^2(1-\eta(X)) + \binom{3}{3}\eta(X)^3\right)\right]$$

$$= E[\eta(X)(1-\eta(X))^3] + 6E[\eta(X)^2(1-\eta(X))^2]$$
$$+ E[\eta(X)^3(1-\eta(X))]$$

Bayes Error Bounds

- Note that $\epsilon_{KNN} = E[\alpha_K(\eta(X))]$, where

$$\alpha_K(p) = p\,P\left(B(k,p) > \frac{k}{2}\right) + (1-p)\,P\left(B(k,p) < \frac{k}{2}\right)$$

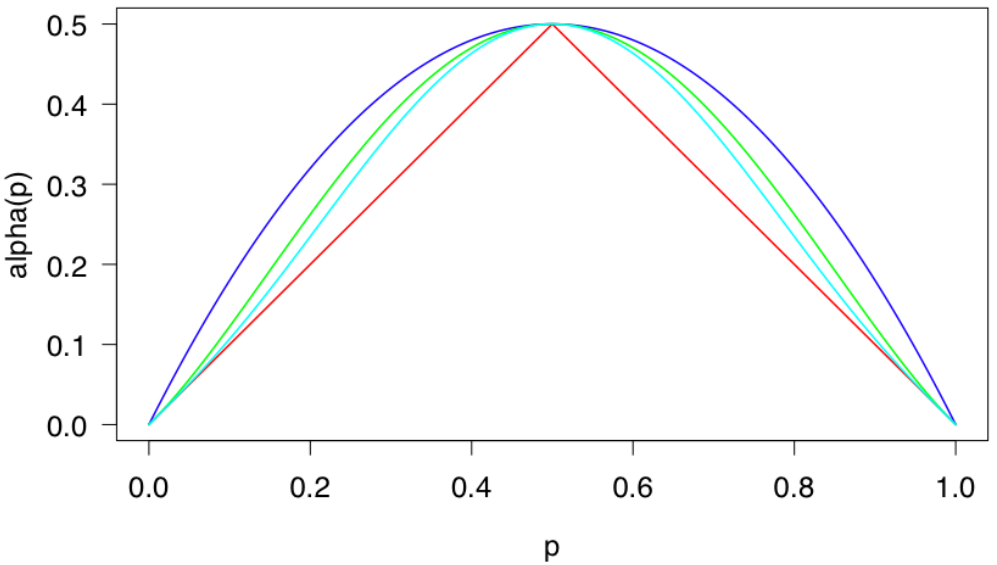
$$= \sum_{i=0}^{(K-1)/2} \binom{K}{i} p^{i+1}(1-p)^{K-i}$$

$$+ \sum_{i=(K+1)/2}^{K} \binom{K}{i} p^{i}(1-p)^{K-i+1}$$

This is a polynomial in p of order $K+1$. The polynomial is concave only for $K = 1$, so *only the 1NN error is an F-error*.

Bayes Error Bounds - II

- The previous fact can be seen in the following plot.



$\min\{\eta, 1-\eta\}$

$\alpha_1(\eta) = 2\eta(1-\eta)$

$\alpha_3(\eta)$

$\alpha_5(\eta)$

- From this plot we have the following inequality

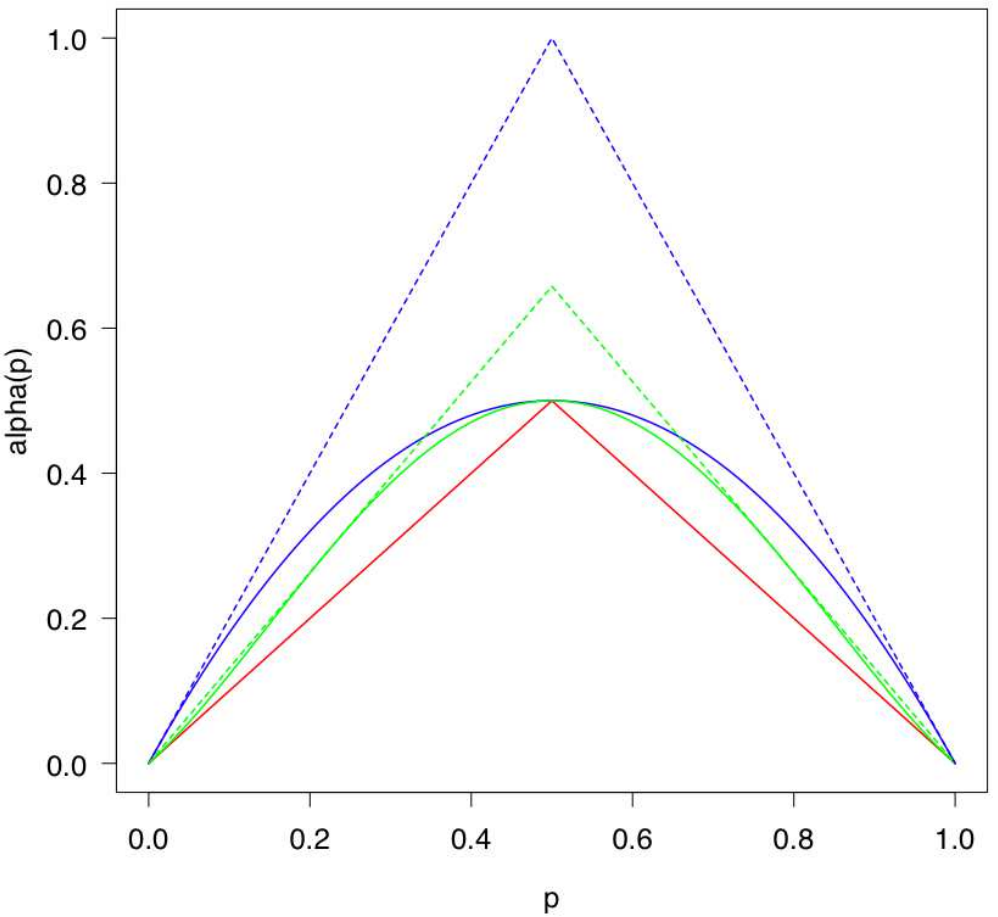
$$\epsilon_{NN} \geq \epsilon_{3NN} \geq \epsilon_{5NN} \geq \cdots \geq \epsilon^*$$

In fact, it can be shown that $\epsilon_{KNN} \rightarrow \epsilon^*$ as $K \rightarrow \infty$.

Bayes Error Bounds - III

- To get bound of the type $\epsilon_{KNN} \leq a_K \epsilon^*$, we let

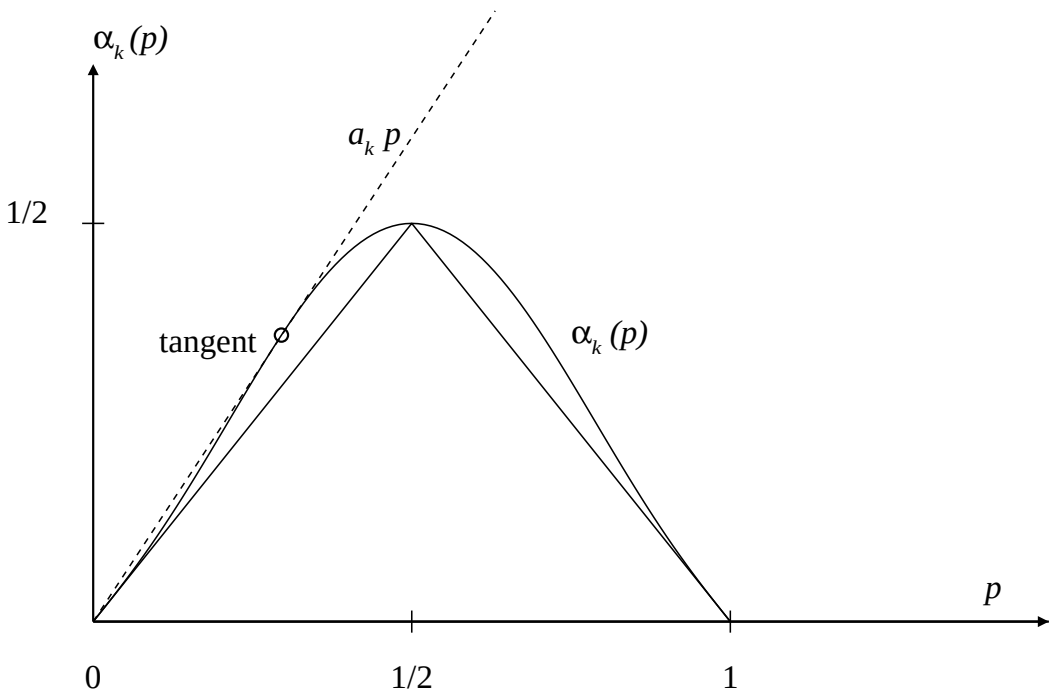
$$a_K = \inf\{ a \mid \alpha_K(p) \leq a \min\{p, 1-p\}\}$$



Bayes Error Bounds - IV

- It can be seen that a_K is the slope of a tangent to $\alpha_K(p)$ that is above $\alpha_K(p)$ and goes through the origin, so that there is a point $0 \leq p_0 \leq \frac{1}{2}$ such that

$$a_K = \alpha'_K(p_0) \text{ where } \alpha'_K(p_0) = \frac{\alpha_K(p_0)}{p_0}$$



Bayes Error Bounds - V

- For example, with $K = 1$,

$$\alpha_1'(p_0) = 2 - 4p_0 = \frac{2p_0(1-p_0)}{p_0} = \frac{\alpha_1(p_0)}{p_0} \Rightarrow p_0 = 0$$

which leads to the familiar bound

$$a_1 = \alpha_1'(0) = 2 \Rightarrow \epsilon_{NN} \leq 2\epsilon^*$$

- For $K = 3$, carrying out the same process yields

$$(p_0)^3 - \frac{4}{3}(p_0)^2 + \frac{1}{4}p_0 = 0$$

Bayes Error Bounds - VI

- This has three solutions:

$$p_0^1 = 0$$

$$p_0^2 = \frac{4+\sqrt{7}}{6}$$

$$p_0^3 = \frac{4-\sqrt{7}}{6}$$

The first is invalid because the tangent is not above $\alpha_K(p)$, the second is invalid because $p_0^2 > 1$, while the last is valid and gives:

$$a_3 = \alpha_3'\left(\frac{4-\sqrt{7}}{6}\right) = \frac{17+7\sqrt{7}}{27} \approx 1.3156 \Rightarrow \epsilon_{3NN} \leq 1.316\epsilon^*$$

which is better than the ϵ_{NN} bound.

Additional Properties

- (DGL Theorem 5.6) For odd K and all distributions F_{XY}

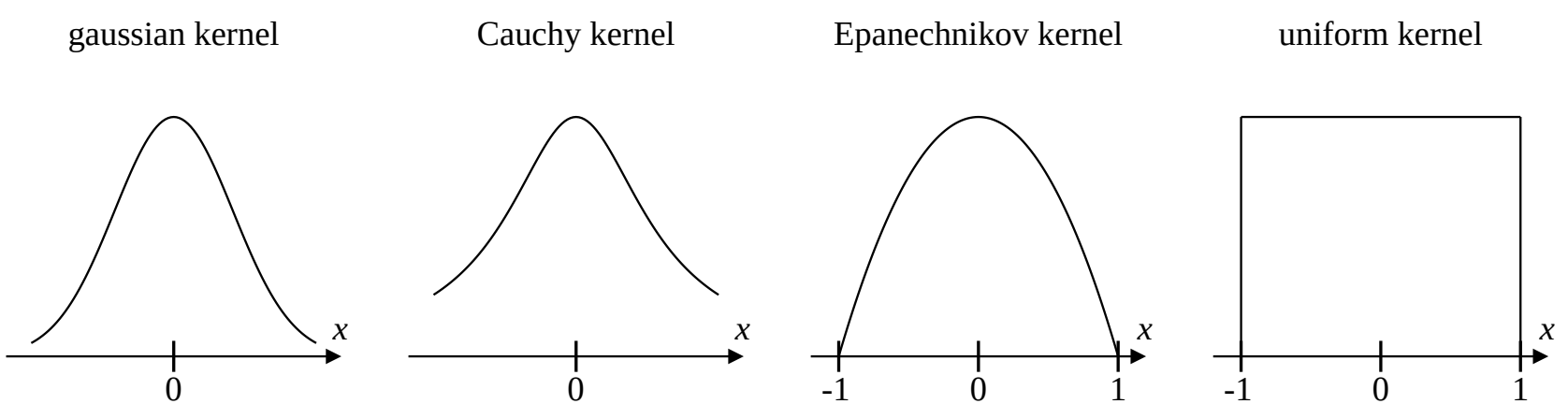
$$\epsilon_{KNN} \leq \epsilon^* + \frac{1}{\sqrt{Ke}}$$

- (DGL Theorem 5.7) For odd K and all distributions F_{XY}

$$\epsilon_{KNN} \leq \epsilon^* + \sqrt{\frac{2\,\epsilon_{NN}}{K}}$$

Kernel Rules

- Kernel rules can be seen as “smoothed” histogram rules, where the cells are allowed to “move” and (or) cell transition is not sharp (so that the weight attributed to points far from the “cell center” decreases smoothly).
- A *kernel* is a function $K : R^d \rightarrow R$ which is usually (but not necessarily, more on this later) nonnegative, and monotonically decreasing on rays starting at the origin.



Examples of Kernels

- The 1-D kernels on the previous figure are given generally in R^d by
  - Gaussian kernel: $K(x) = e^{-||x||^2}$
  - Cauchy kernel: $K(x) = \frac{1}{1+||x||^{d+1}}$
  - Epanechnikov: $K(x) = (1 - ||x||^2)\, I_{\{||x||\leq 1\}}$
  - Uniform (spherical): $K(x) = I_{\{||x||\leq 1\}}$
  - Uniform (cubic):

$$K(x) = \begin{cases} 1\,, & |x_j| \leq \frac{1}{2}\,, \text{for all } j = 1, \ldots, d \\ 0\,, & \text{otherwise} \end{cases}$$

Kernel Classifier

- Provided $K(x) > 0$, for all $x \in R^d$, the estimate $\eta_n(x)$ of $\eta(x)$ is derived as:

$$\eta_n(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right) I_{\{Y_i=1\}}}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)}$$

- The designed classifier is then given by:

$$\psi_n(x) = \begin{cases} 1, & \eta_n(x) > \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

$$= \begin{cases} 1, & \sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right) I_{\{Y_i=1\}} \geq \sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right) I_{\{Y_i=0\}} \\ 0, & \text{otherwise} \end{cases}$$

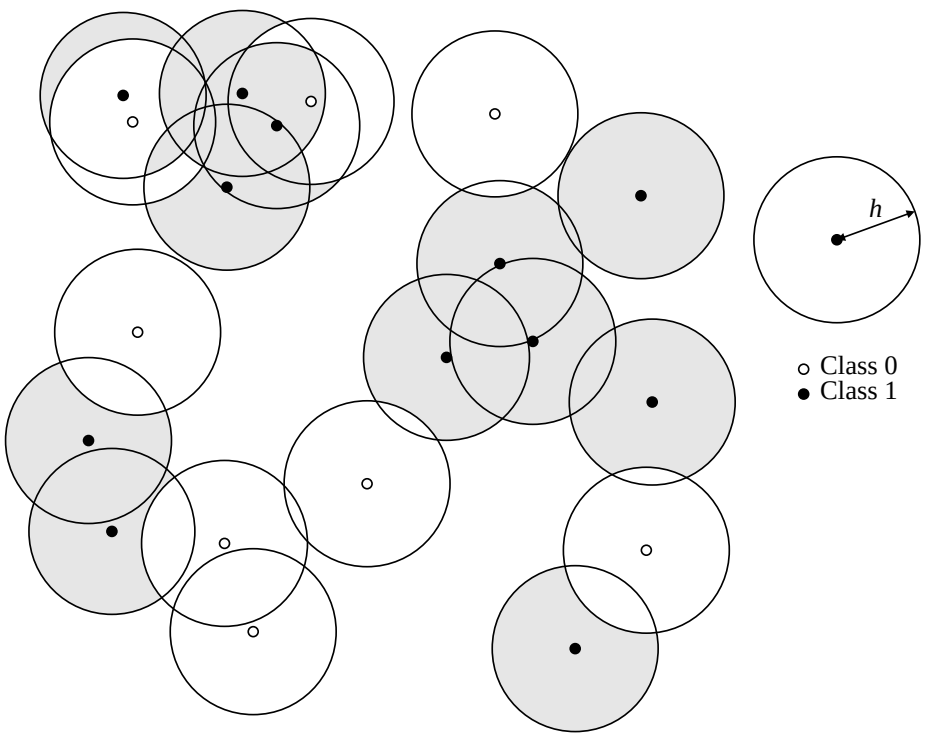
Kernel Classifier - II

- This can be seen as adding the distance-weighted "influences" of each data point (X_i, Y_i) on x and assigning the label of the most "influent" class.

- Note in the previous slide that $\eta_n(x)$ needs the condition $K > 0$ to make sure it is a valid probability, but the expression for $\psi_n(x)$ *does not* need this assumption, that is, the kernel can take negative values and still define a valid classifier. This indicates that pattern recognition is more general than pure density estimation (more on this later).

Moving-Window Rule

- In the case of a uniform kernel, the classifier is sometimes called a “moving window” classifier. It can be seen as a histogram rule where the cells are allowed to move.



Amount of Kernel Smoothing

- The smoothing parameter here is h_n.
- If h_n is too small, the rule is too "local" (only the closest points exert influence on a given x) which leads to more overfitting.
- If h_n is too large, the rule is too "global" (far points exert influence on a given x) which leads to too much smoothing.
- In both cases, the true classification error increases. How to pick the optimal h_n is a model selection issue, which will be addressed later in the class.

PR is not Density Estimation

- The statement that one should estimate densities to find a good classifier is problematic, because good density estimation requires more data than good classifier design.
- There is another reason however why Pattern Recognition is different than density estimation, and this is illustrated by the Kernel rule.
- Density estimation is given by

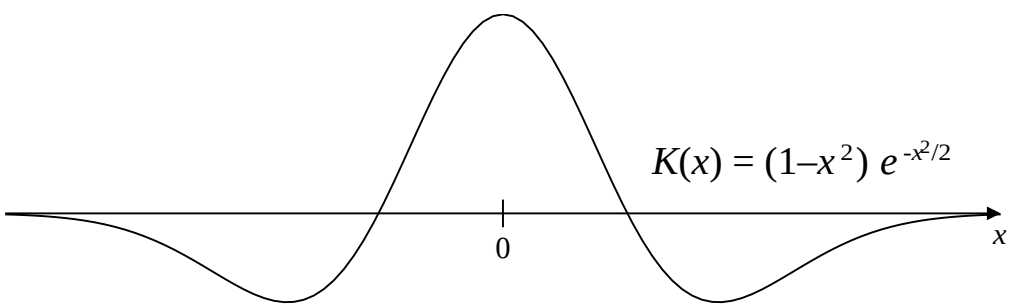
$$p_n(x) = \frac{1}{nh_n^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)$$

where one *must* have $K \geq 0$ and $\int K(u)\, du = 1$ in order to guarantee $p_n(x)$ is always a probability density.

PR is not Density Estimation - II

- In Pattern Recognition (more specifically, in kernel classification), this is not so. For example, an example of kernel that can be used for PR but not for density estimation is the *hermite* kernel. Note that here $K \ngeq 0$.

$$K(x) = (1 - ||x||^2)\, e^{-||x||^2}$$



- Therefore, density estimation methods such as Parzen Windows may be difficult to apply with small samples and are also not general enough for PR.

Universal Consistency

- A kernel is called *regular* if it is nonnegative ($K \geq 0$), bounded away from zero around a neighborhood of the origin, bounded, uniformly continuous, and integrable. This is a technical definition, but it suffices to have in mind that all example of kernels discussed in this lecture are regular, with the exception of the Hermite kernel, because it is not nonnegative.

- (DGL Therem 10.1) If X has a density, the kernel K is regular, and $h_n \to 0$ with $nh_n^d \to \infty$ as $n \to \infty$, then $\epsilon_n \to \epsilon^*$ with probability 1 for all distributions F_{XY}, that is, the kernel rule is strongly universally consistent.