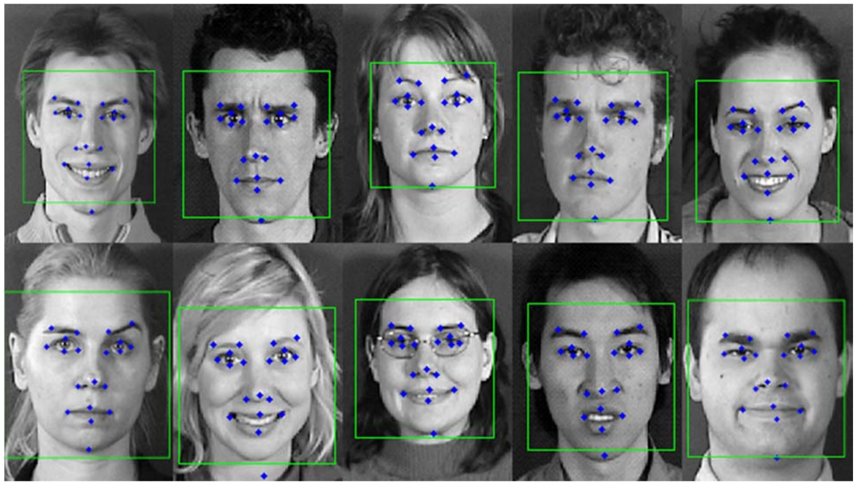
Learning from Data
4. Classification: Nearest Neighbor

Spring 2017

Classification

- Supervised (preditive) learning: given examples with labels, predict labels for all unseen examples
  - Classification:
    - label = category,
    - $\mathbf{y} \in \mathcal{Y} = \{1, \ldots, m\}$, m = number of classes
    - $\ell(\mathbf{x}, y, h) = 1(h(\mathbf{x}) \neq y)$, Risk = $P(Y \neq h(\mathbf{X})) = P(\mathrm{Error})$



$\mathbf{x}$ = facial geometry features

y = gender label

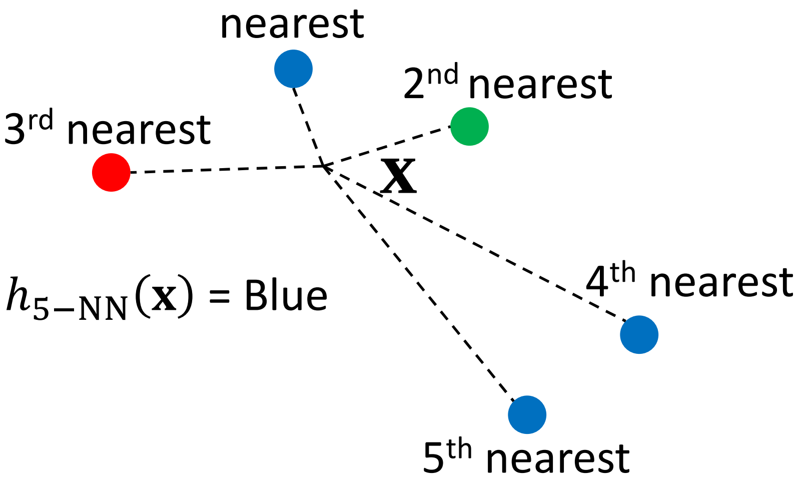
k-Nearest Neighbor (NN) Classifier

- Discriminative classifier: only $p(y|\mathbf{x})$ estimated
- Non-parametric: no parametric model for $p(y|\mathbf{x})$
- Example of memory-based learning
- Need 2 ingredients to specify classifier:
  1. k : number of nearest neighbors, typically chosen to be not a multiple of m (the number of classes)
  2. $\mathrm{dist}: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$, a measure of “nearness” $\mathrm{dist}(\mathbf{x}, \mathbf{x}')$ between $\mathbf{x}$ and $\mathbf{x}$’

  – typically, dist is chosen to be a metric, i.e., it is positive, definite, i.e., = 0 if, and only if, $\mathbf{x} = \mathbf{x}'$, symmetric, i.e., $\mathrm{dist}(\mathbf{x}, \mathbf{x}') = \mathrm{dist}(\mathbf{x}', \mathbf{x})$, and satisfies the triangle inequality, i.e., $\mathrm{dist}(\mathbf{x}_1, \mathbf{x}_3) \leq \mathrm{dist}(\mathbf{x}_1, \mathbf{x}_2)$+$\mathrm{dist}(\mathbf{x}_2, \mathbf{x}_3)$ for any three points

k-Nearest Neighbor Classifier

- Classifier description in words: $h_{k-\mathrm{NN}}(\mathbf{x}) =$ most abundant label (majority vote) among the labels of the k nearest training examples of $\mathbf{x}$ (breaking ties in some way)



$h_{5-\mathrm{NN}}(\mathbf{x})$ = Blue

k-Nearest Neighbor Classifier

- Formal description: Given labeled training data $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ and a test point $\mathbf{x}$

- Let $\{(\mathbf{x}_{(1)}, y_{(1)}), \ldots, (\mathbf{x}_{(n)}, y_{(n)})\}$ be a re-ordering of training data such that
$$d(\mathbf{x}, \mathbf{x}_{(1)}) \leq d(\mathbf{x}, \mathbf{x}_{(2)}), \ldots, d(\mathbf{x}, \mathbf{x}_{(n)})$$

- Then,
$$h_{k-\mathrm{NN}}(\mathbf{x}) = \arg \max_{y=1,\ldots,m} \underbrace{\sum_{j=1}^{k} 1\left(y_{(j)} = y\right)}_{\text{number of } k \text{ NNs of } \mathbf{x} \text{ with label } = y}$$

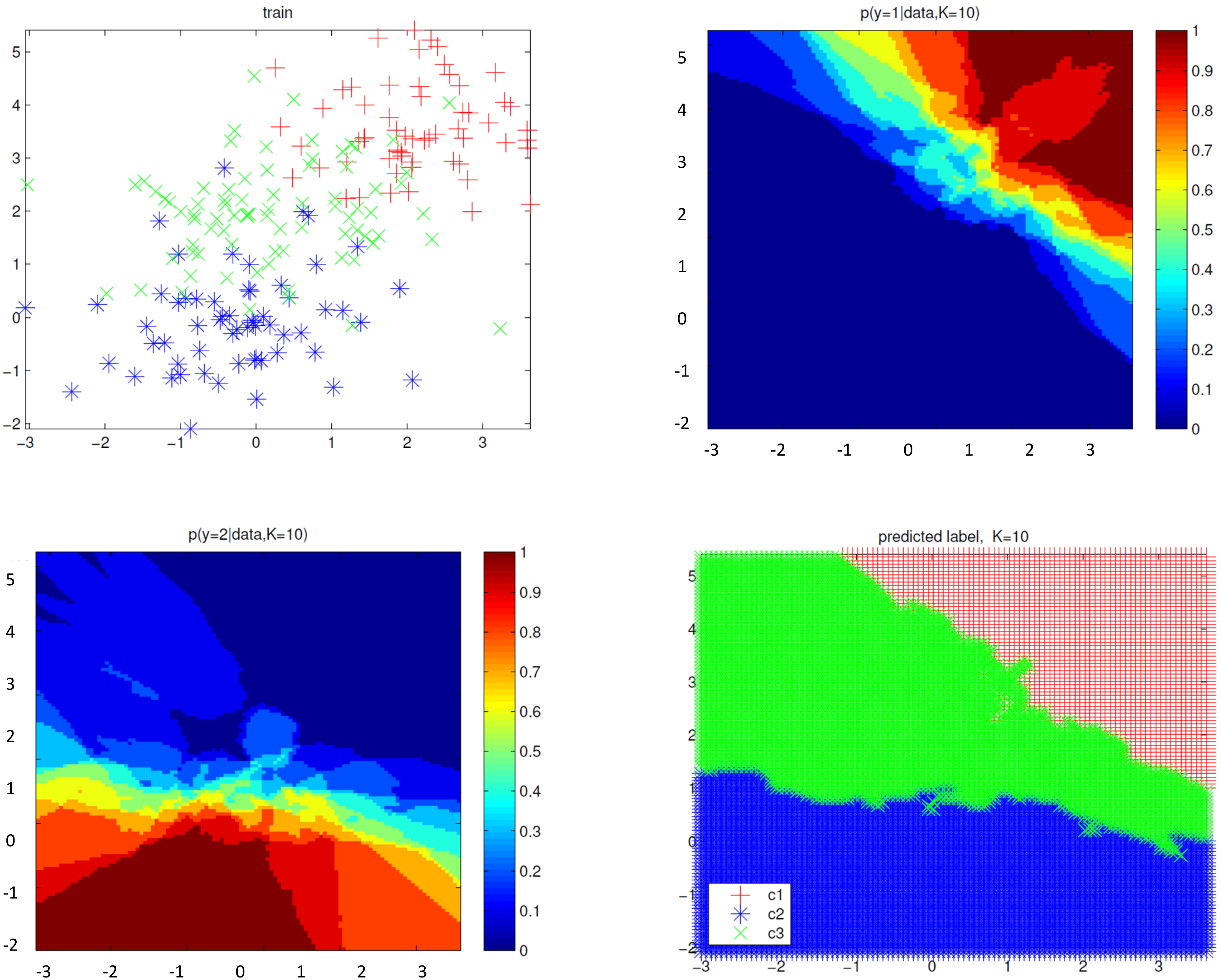
k-Nearest Neighbor Classifier

- Discriminative model based interpretation:

$$\widehat{p}_{k-\mathrm{NN}}(y|\mathbf{x}) = \underbrace{\frac{1}{k}\sum_{j=1}^{k} 1\big(y_{(j)} = y\big)}_{\text{fraction of } k \text{ NNs of } \mathbf{x} \text{ with label } = y}$$

- This is a non-parametric estimate of $p(y|\mathbf{x})$, since the number of parameters = n grows with training data

- k-NN classifier = MPE/MAP rule with the above non-parametric estimate of $p(y|\mathbf{x})$.



train

p(y=1|data,K=10)

p(y=2|data,K=10)

predicted label, K=10

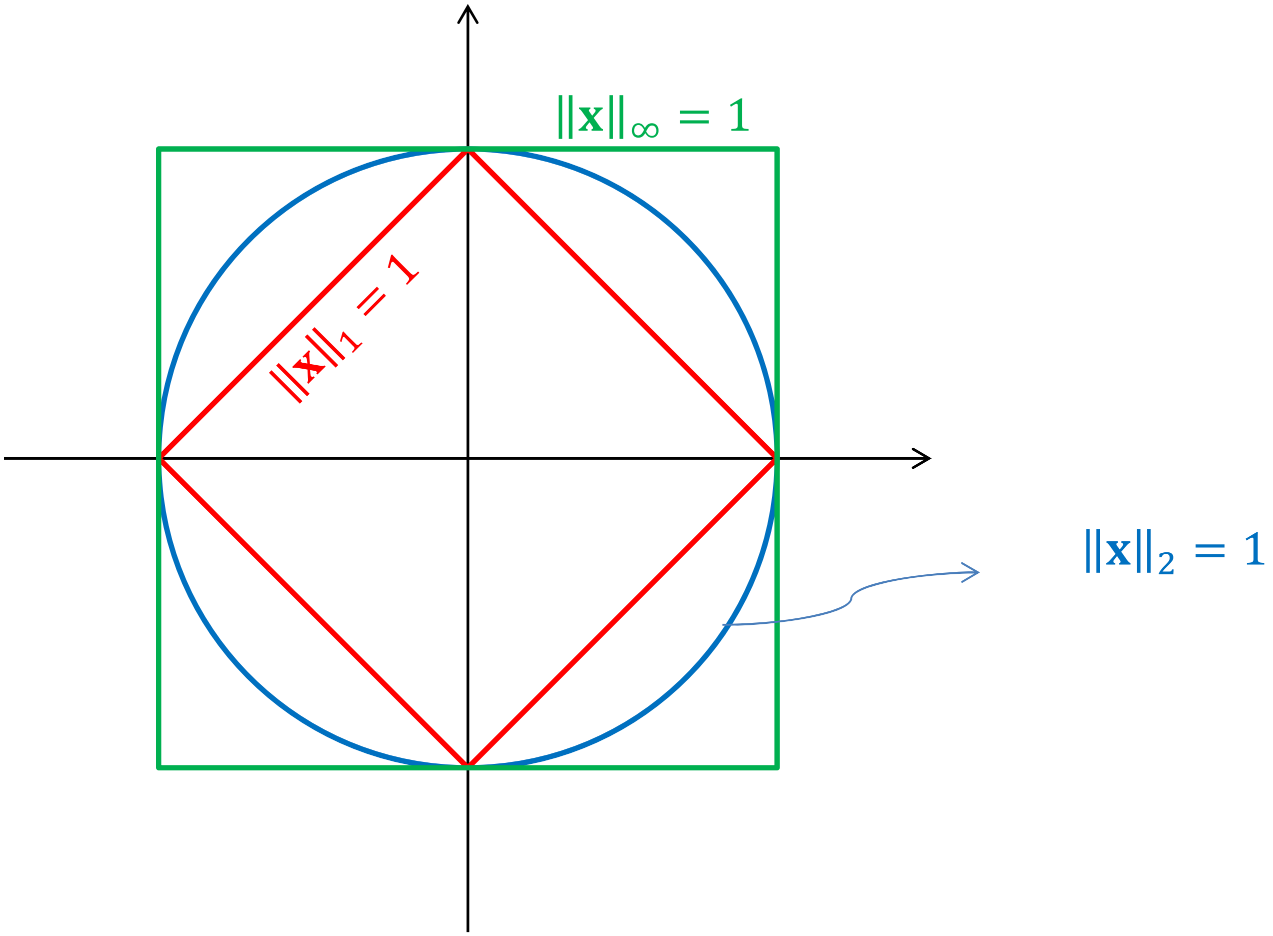
Common distance functions

- ℓ_p distance:

$$\|\mathbf{x}-\mathbf{x}'\|_p := \left(\sum_{i=1}^{d} |x_i - x_i'|^p\right)^{\frac{1}{p}}$$

  - can prove that this is a norm-distance for all $p \geq 1$
- Euclidean or ℓ_2 distance: ℓ_p distance with $p = 2$
- taxi-cab or city-block or Manhattan or ℓ_1 distance: ℓ_p distance with $p = 1$
- max norm or ℓ_∞ distance: limit of ℓ_p as $p \to \infty$. Can be shown to be equal to: $\max_{1 \leq i \leq d} |x_i - x_i'|$
- Mahalanobis distance: $\sqrt[2]{(\mathbf{x}-\mathbf{x}')^T \Sigma^{-1} (\mathbf{x}-\mathbf{x}')}$ with Σ a positive definite square matrix

Common distance functions



Contours of constant p-norms
(distance from origin)
in 2 dimensions (d=2) for different p

Voronoi regions for 1-NN classifiers

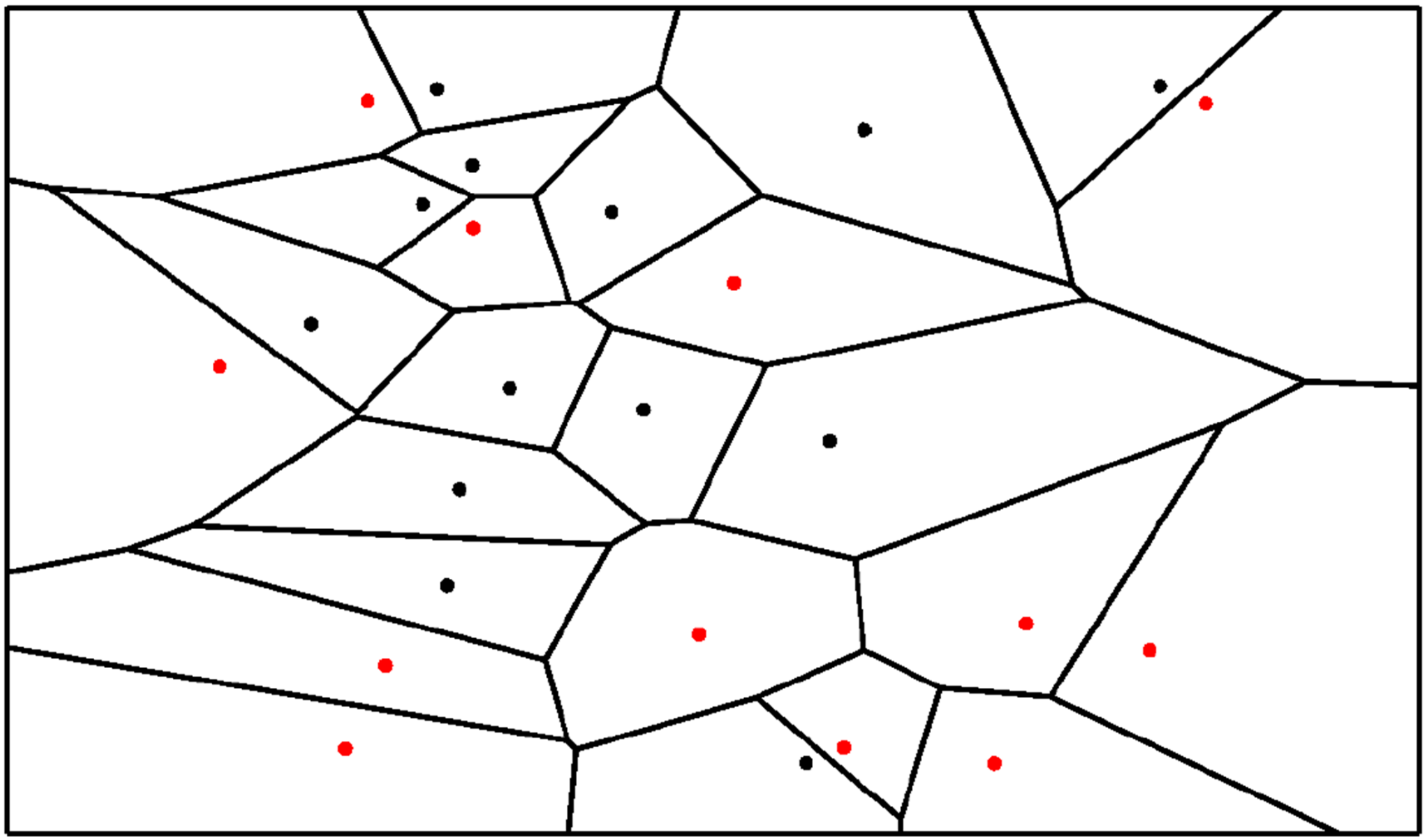
- Voronoi region of a training point $\mathbf{x}_j$ = all points in feature space that are closer to $\mathbf{x}_j$ than to any other training point:

$$\mathcal{V}(\mathbf{x}_j) := \{\mathbf{x} : d(\mathbf{x}, \mathbf{x}_j) \leq d(\mathbf{x}, \mathbf{x}_{j'}), \forall j' \neq j\}$$

- ⇒The 1-NN classifier will assign all points in $\mathcal{V}(\mathbf{x}_j)$ the same class label as that of $\mathbf{x}_j$
- The Voronoi tessellation is a partition of the feature space into the Voronoi regions of the training points
- If the distance function is a norm induced by an inner product, e.g.,$\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$, then the Voronoi regions are convex, i.e., for any two points in the region, the entire line segment joining them is also entirely in the region.

Voronoi regions for 1-NN classifiers

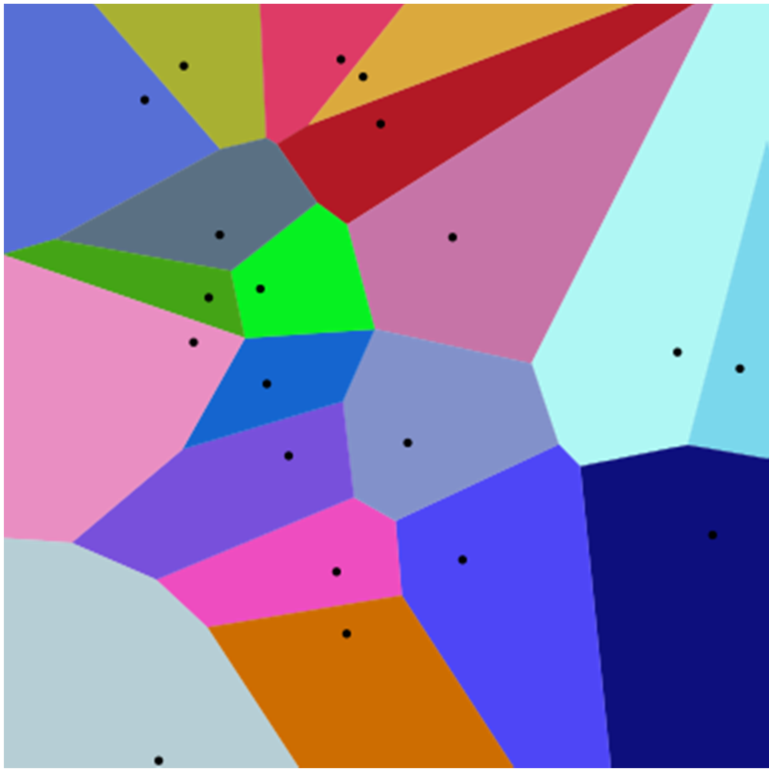
- Voronoi tessellation of the 2-dim plane induced by the 1-NN classifier based on Euclidean distance

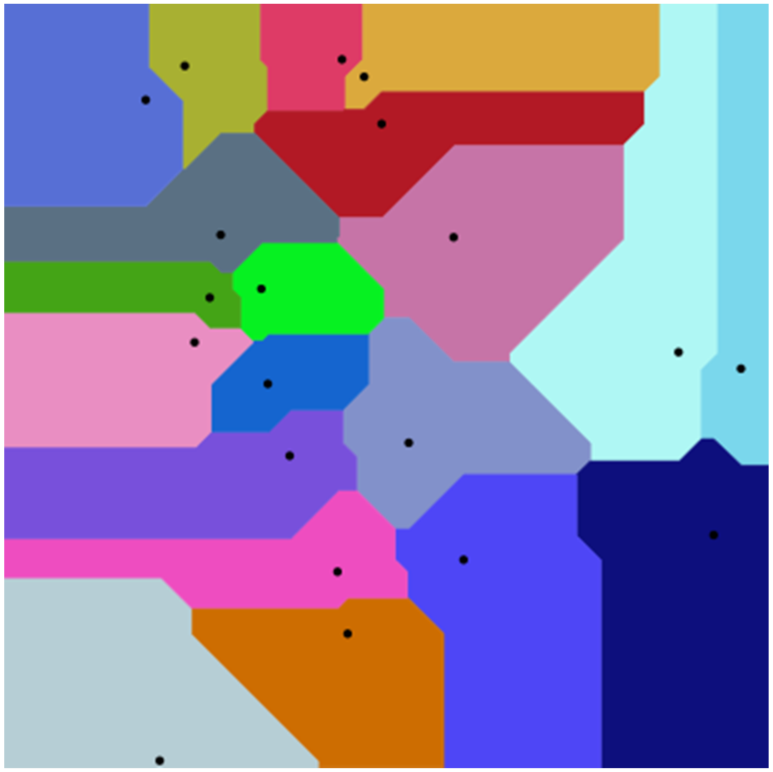


Voronoi regions for 1-NN classifiers

- Voronoi tessellation based on:

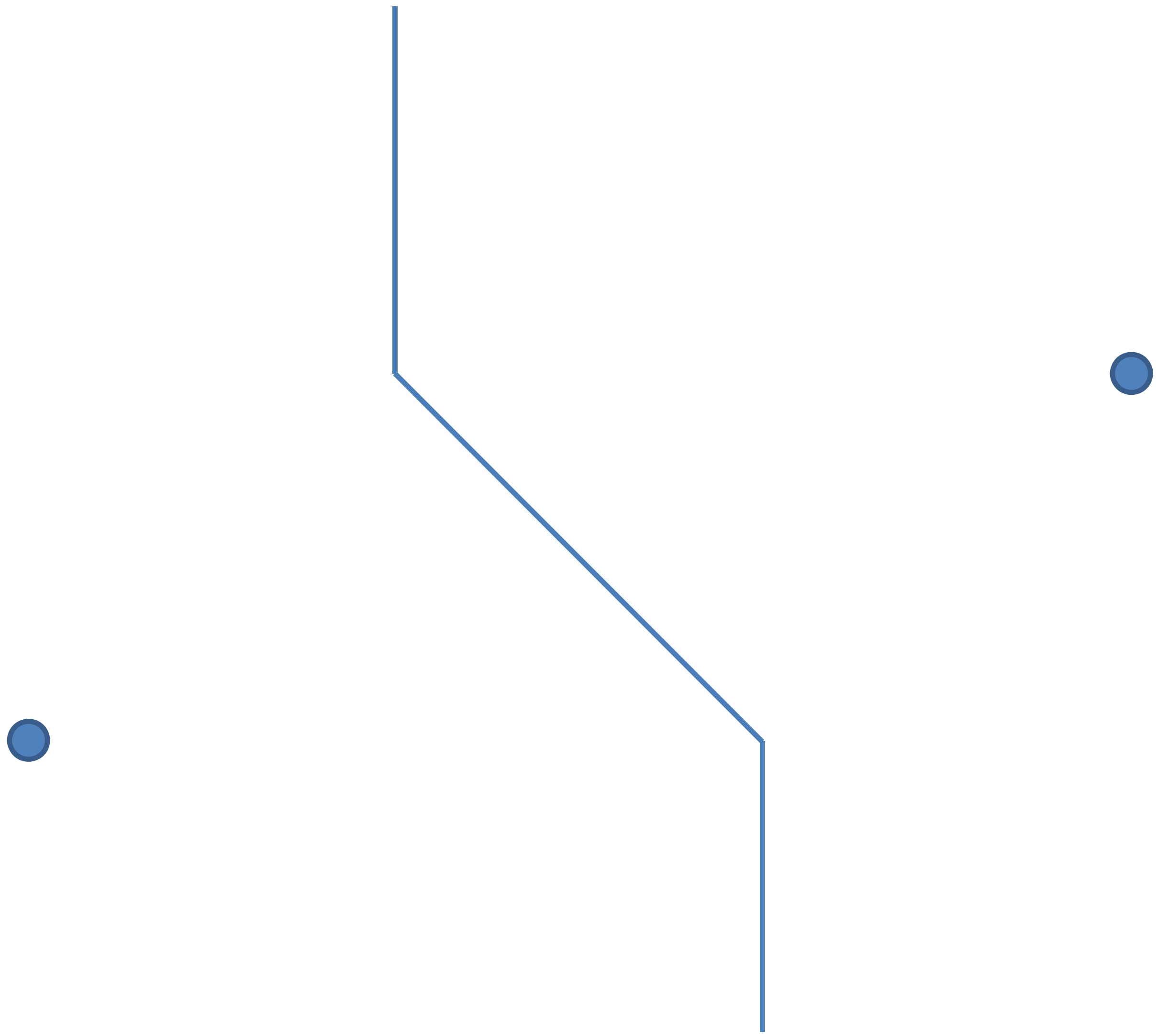
Euclidean distance (convex) versus Manhattan distance (non-convex)





Voronoi regions for 1-NN classifiers

- 2-point Voronoi diagram for Manhattan distance



Remarks

- Asymptotic performance guarantees:
  - as $n \to \infty$, risk of 1-NN classifier (for 0-1 loss) $\leq 2R_{\text{Bayes}}$ for all “nice” data distributions and distance functions!
  - If as $n \to \infty$,
    - $k_n \to \infty$ …(ensures more training examples in majority vote)
    - $\frac{k_n}{n} \to 0$ ... (ensures the k_n NNs get closer to any test point)

    then, risk of k_n–NN classifier (for 0-1) loss $\to R_{\text{Bayes}}$ for all “nice” data distributions and distance functions!!

- In practice, the best value of k is determined via cross-validation

Risk of 1-NN classifier (for 0-1 loss) $\leq 2R_{\text{Bayes}}$

Result: as $n \to \infty$,

$$P(Y_{\text{test}} \neq h_{1-NN}(\mathbf{X}_{\text{test}})) \leq 2P(Y_{\text{test}} \neq h_{\text{Bayes}}(\mathbf{X}_{\text{test}}))$$

for all "nice" joint distributions $p(y|\mathbf{x})p(\mathbf{x})$ and distance functions $\text{dist}(\cdot,\cdot)$

Intuition for m = 2 classes. Let $\mathcal{Y} = \{0,1\}$.

- Training set: $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n) \sim \text{IID}\ p(y|\mathbf{x})p(\mathbf{x})$
- Test point: $(\mathbf{X}_{\text{test}}, Y_{\text{test}}) \sim p(y|\mathbf{x})p(\mathbf{x})$ independent of training set
- Let $NN(\mathbf{x})$ = training feature closest to $\mathbf{x}$ and let $Y_{NN(\mathbf{x})}$ denote its label

Risk of 1-NN classifier (for 0-1 loss) $\leq 2R_{\text{Bayes}}$

- Key observation 1: if $\mathbf{x}$ is within the interior of the support of $p(\mathbf{x})$ and $p(\mathbf{x})$, $\text{dist}(\cdot,\cdot)$ are "nice", then as $n \to \infty, NN(\mathbf{x}) \to \mathbf{x}$ almost surely:

$$\forall \epsilon > 0, P(\text{dist}(\mathbf{x}, NN(\mathbf{x}) \geq \epsilon)) = (1 - \underbrace{P(\text{dist}(\mathbf{x}, \mathbf{X}) \geq \epsilon))}_{>0 \text{ since } \mathbf{x} \in \text{int supp}(p(\mathbf{x})) \text{ and } p(\mathbf{x}), \text{dist}(\cdot,\cdot) \text{ are nice}}{}^n \to 0$$

- Key observation 2: if $p(y|\mathbf{x})$ is "nice" then whenever $\mathbf{x}'$ is close to is $\mathbf{x}$ (as measured by $\text{dist}(\cdot,\cdot)$) then $p(y|\mathbf{x}')$ will be close to $p(y|\mathbf{x})$

- Key observation 3:

$$P(Y_{\text{test}} \neq h_{\text{Bayes}}(\mathbf{X}_{\text{test}})|\mathbf{X}_{\text{test}} = \mathbf{x}) = \min\{p(0|\mathbf{x}), p(1|\mathbf{x})\}$$

since $h_{Bayes}(\mathbf{x}) = \arg\max_{y=0,1} p(y|\mathbf{x})$

Risk of 1-NN classifier (for 0-1 loss) $\leq 2R_{\text{Bayes}}$

- Proof-sketch:

$$
\begin{aligned}
&P(Y_{\text{test}} \neq h_{1-NN}(\mathbf{X}_{\text{test}})|\mathbf{X}_{\text{test}} = \mathbf{x}) \\
&= P(Y_{\text{test}} \neq Y_{NN(\mathbf{x})}|\mathbf{X}_{\text{test}} = \mathbf{x}) \\
&= P(Y_{\text{test}} = 1|\mathbf{X}_{\text{test}} = \mathbf{x})P(Y_{NN(\mathbf{x})} = 0|\mathbf{X}_{\text{test}} = \mathbf{x}) \\
&\quad + P(Y_{\text{test}} = 0|\mathbf{X}_{\text{test}} = \mathbf{x})P(Y_{NN(\mathbf{x})} = 1|\mathbf{X}_{\text{test}} = \mathbf{x}) \\
&= p(1|\mathbf{x})P(Y_{NN(\mathbf{x})} = 0|\mathbf{X}_{\text{test}} = \mathbf{x}) + p(0|\mathbf{x})P(Y_{NN(\mathbf{x})} = 1|\mathbf{X}_{\text{test}} = \mathbf{x}) \\
&\approx p(1|\mathbf{x})P(Y_{\text{test}} = 0|\mathbf{X}_{\text{test}} = \mathbf{x}) + p(0|\mathbf{x})P(Y_{\text{test}} = 1|\mathbf{X}_{\text{test}} = \mathbf{x}) \\
&= p(1|\mathbf{x})p(0|\mathbf{x}) + p(0|\mathbf{x})p(1|\mathbf{x}) \\
&= 2p(0|\mathbf{x})p(1|\mathbf{x}) \\
&\leq 2\min\{p(0|\mathbf{x}), p(1|\mathbf{x})\} \\
&= 2P(Y_{\text{test}} \neq h_{\text{Bayes}}(\mathbf{X}_{\text{test}})|\mathbf{X}_{\text{test}} = \mathbf{x})
\end{aligned}
$$

training and test samples are independent

Key observations 1 and 2

If $a, b \in [0,1] \Rightarrow ab \leq a$ and $ab \leq b \Rightarrow ab \leq \min\{a, b\}$

Key observation 3

- The result follows by taking expectation over $\mathbf{X}_{\text{test}}$

Remarks

- Useful observation: Euclidean distances can be computed via inner products: $\|\mathbf{x}-\mathbf{x}'\|_2^2 = \langle \mathbf{x},\mathbf{x}\rangle + \langle \mathbf{x}',\mathbf{x}'\rangle - 2\langle \mathbf{x},\mathbf{x}'\rangle$
- Let $\mathbb{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$, $\mathrm{diag}(A)$ = column vector of the main diagonal elements of square matrix A, and $\mathbf{1}$ = column vector of all ones. Then the square Euclidean distance matrix (EDM) is given by:

$$\mathrm{EDM}(\mathbb{X}) = \mathbf{1} * \mathrm{diag}\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{\mathrm{T}} - 2\mathbb{X}^{\mathrm{T}}\mathbb{X} + \mathrm{diag}\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right) * \mathbf{1}^{\mathrm{T}}$$

$$\mathrm{EDM}(\mathbb{X})(i,j) = \left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2^2$$

- Note: $\mathrm{rank}\left(\mathrm{EDM}(\mathbb{X})\right) \leq d + 2$

Weighted nearest neighbor rules

- Observation: $$\begin{aligned} h_{k-NN}(\mathbf{x}) &= \arg\max_y \widehat{p}_{k-\mathrm{NN}}(y|\mathbf{x}) \\ &= \arg\max_y \sum_{j=1}^{k} \frac{1}{k} 1\big(y_{(j)} = y\big) \\ &= \arg\max_y \sum_{j=1}^{n} \underbrace{w(\mathbf{x}, \mathbf{x}_j)}_{\text{weights}} 1(y_j = y) \end{aligned}$$

where

$$w(\mathbf{x}, \mathbf{x}_j) = \begin{cases} \frac{1}{k} & \text{if } \mathbf{x}_j \in \{k\text{NN's of } \mathbf{x}\} \\ 0 & \text{otherwise} \end{cases}$$

- Idea: use alternative weight function $w(\mathbf{x}, \mathbf{x}')$: nonnegative and decreases with increasing distance $\mathrm{dist}(\mathbf{x}, \mathbf{x}')$. Can be interpreted as a "similarity" score. Typically,

$$w(\mathbf{x}, \mathbf{x}') \propto e^{-\mathrm{dist}(\mathbf{x}, \mathbf{x}')}$$

Weighted nearest neighbor rules

Example:

$$w(\mathbf{x},\mathbf{x}') = \frac{1}{\alpha\sqrt{2\pi}} e^{-\frac{1}{2\alpha^2}\|\mathbf{x}-\mathbf{x}'\|^2} = \frac{1}{\alpha}\psi\left(\frac{\|\mathbf{x}-\mathbf{x}'\|}{\alpha}\right), \psi(t) = \mathcal{N}(0,1)(t)$$

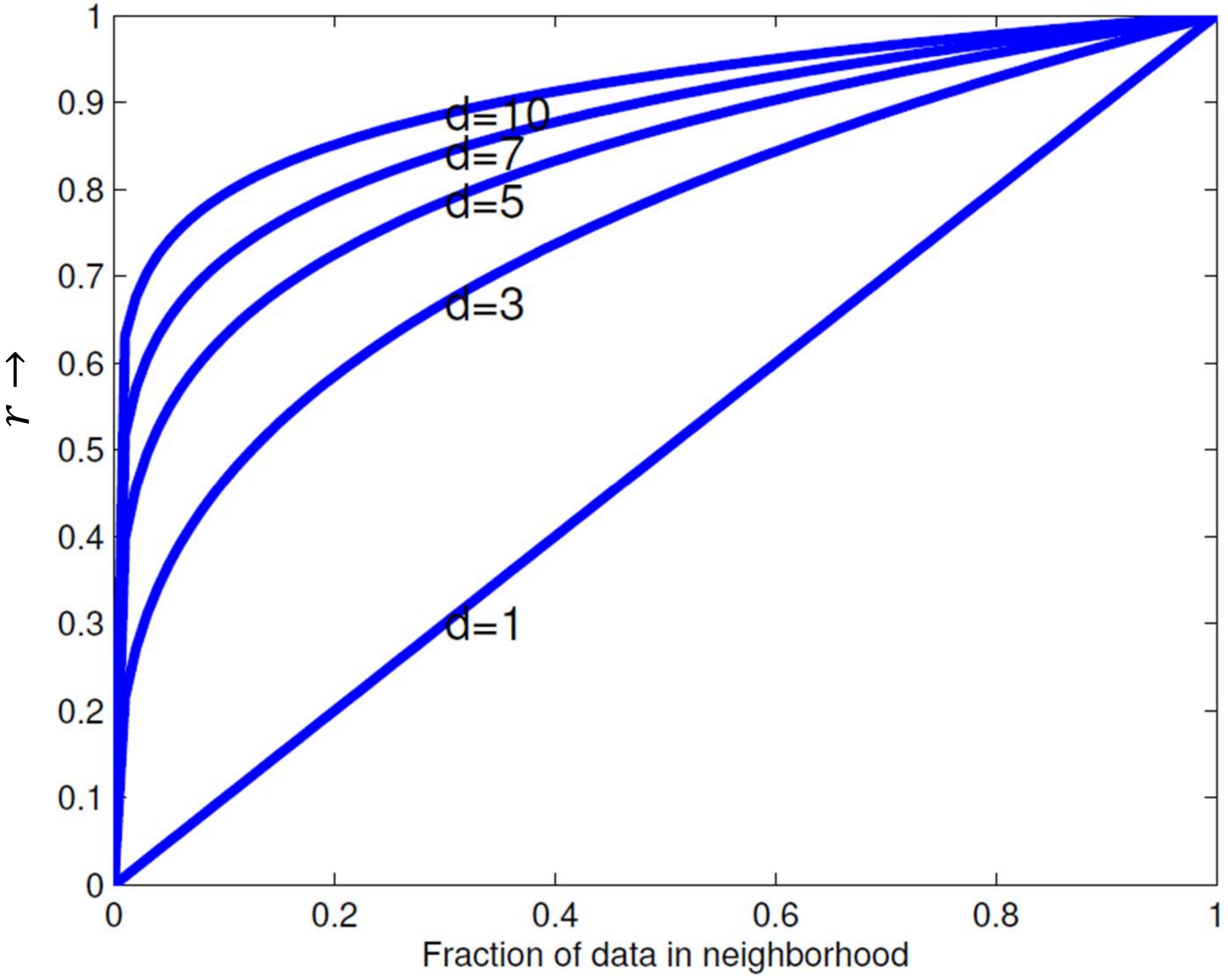
Here, α is a bandwidth parameter which controls the neighborhood size like k in k-NN (small α ⇒small neighborhood)

- $$\widehat{p}(\mathbf{x}|y) = \frac{1}{n_y}\sum_{j=1}^{n} \frac{1}{\alpha\sqrt{2\pi}} e^{-\frac{1}{2\alpha^2}||\mathbf{x}-\mathbf{x}_j||^2} 1(y_j = y), n_y = |\{j : y_j = y\}|$$
- is a non-parametric estimate of the conditional density $p(\mathbf{x}|y)$
- called kernel-density estimate with the Gaussian “kernel” of bandwidth α
- Thus, $$\widehat{p}(y|\mathbf{x}) \propto \widehat{p}(\mathbf{x}|y)\underbrace{\widehat{p}(y)}_{\frac{n_y}{n}} = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{\alpha\sqrt{2\pi}} e^{-\frac{1}{2\alpha^2}||\mathbf{x}-\mathbf{x}_j||^2} 1(y_j = y)$$
- The weighted nearest neighbor rule can therefore be interpreted as implementing the MPE rule (= MAP rule) with the above non-parametric estimate of the joint distribution $p(\mathbf{x}, y)$
- Other kernel fns.: $\psi(t)$ any “nice” symmetric pdf, nonincreasing in $|t|$

Curse of dimensionality

- NN classifiers break down if d is large
- Suppose training points uniformly distributed in d-dimensional unit sphere centered at origin
- To "capture" a fraction f of all training points, the distance r from the origin that one needs to travel is given by: $\frac{r^d}{1^d} = f \Rightarrow r = f^{\frac{1}{d}}$
- Say $d = 10$ (modest).
  - $f = 0.1 \Rightarrow r = 0.8$, i.e., to capture 10% of samples, need to traverse 80% of range! Samples that are so far from a point (non-local) are typically not good predictors of behavior at that point
  - $f = 0.01 \Rightarrow r = 0.63$, i.e., to capture only 1% of samples, still need to traverse 63% of range!
  - Seen another way, to keep decision making local, if we fix $r = 0.1 \Rightarrow$ number of points used for classification = $nr^d = n10^{-10}$ $\Rightarrow$ will need a HUMONGOUS amount of training data to make a reliable decision

Curse of dimensionality



- Turns out that in high dimensions, most points are far from each other!
- Also turns out that in high dimensions, most points are almost orthogonal to each other!! …Weird things happen in high dimensions