8. Decisions and Bayes classifier

- Statistical decision theory
- Bayes classifier
- K nearest neighbors (KNN)

Many slides borrowed from Prof. Minh Hoai Nguyen's previous course offering

Statistical decision theory

- A game against nature
- Using observations, make a choice
- Maximize utility / minimize loss

$$\hat{y} = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, \mathcal{U}(y|x) = \underset{y \in \mathcal{Y}}{\operatorname{argmin}}\, \mathcal{L}(y|x)$$

- Nature doesn't act back (not interactive)

Tasks and loss functions

- Mean squared error: $\mathcal{L}(\hat{\theta}, \theta) = \|\theta - \hat{\theta}\|_2^2$
- Regression error: $\mathcal{L}(x^T\hat{\theta}, y) = \frac{1}{2}\|x^T\hat{\theta} - y\|^2$
- Classification error: $\mathcal{L}(x^T\hat{\theta}, y) = \begin{cases} 1 & \text{if } \mathbf{sign}(x^T\hat{\theta}) \neq y \\ 0 & \text{else.} \end{cases}$
- Density error: $\mathcal{L}(y|\hat{\theta}, y|\theta) = \int p(y|\theta) \log \frac{p(y|\theta)}{p(y|\hat{\theta})} dy$ (KL divergence)

Today's goal With respect to losses, how well did we do?

Today's "advisers"

Two omnipotent advisers

- Minimax estimator
- Bayes estimator

One practical algorithm

- K-nearest neighbors

Minimax estimator

Given

- data $x \in \mathcal{X}$
- true response $y(x) \in \mathcal{Y}$
- estimate $\hat{y} \in \mathcal{Y}$
- loss $\mathcal{L}(y, \hat{y})$

Minimax estimator

Given

- data $x \in \mathcal{X}$
- true response $y(x) \in \mathcal{Y}$
- estimate $\hat{y} \in \mathcal{Y}$
- loss $\mathcal{L}(y, \hat{y})$

Max risk

$$r_{\max}(\hat{y}) = \max_{y \in \mathcal{Y}} \mathcal{L}(\hat{y}; y(x))$$

$\hat{y}_{\text{minimax}}(x)$ is a minimax estimator if

$$\hat{y}_{\text{minimax}}(x) = \operatorname*{argmin}_{\hat{y} \in \mathcal{Y}} r_{\max}(\hat{y})$$

Minimax estimator

Given

- data $x \in \mathcal{X}$
- true response $y(x) \in \mathcal{Y}$
- estimate $\hat{y} \in \mathcal{Y}$
- loss $\mathcal{L}(y, \hat{y})$

Max risk

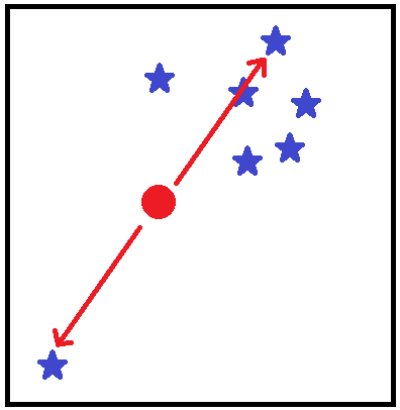
$$r_{\max}(\hat{y}) = \max_{y \in \mathcal{Y}} \mathcal{L}(\hat{y}; y(x))$$

$\hat{y}_{\text{minimax}}(x)$ is a minimax estimator if

$$\hat{y}_{\text{minimax}}(x) = \operatorname*{argmin}_{\hat{y} \in \mathcal{Y}} r_{\max}(\hat{y})$$

Example: point estimation $(y(x) = x)$

- $\mathcal{Y} = \{\star\text{'s}\}$, $\hat{y}_{\text{minimax}} =$ ●
- Minimax estimator is completely defined by a few "worst players"



Bayes estimator

Given

- data $x \in \mathcal{X}$
- true response $y(x) \in \mathcal{Y}$
- estimate $\hat{y} \in \mathcal{Y}$
- loss $\mathcal{L}(y, \hat{y})$

Bayes risk

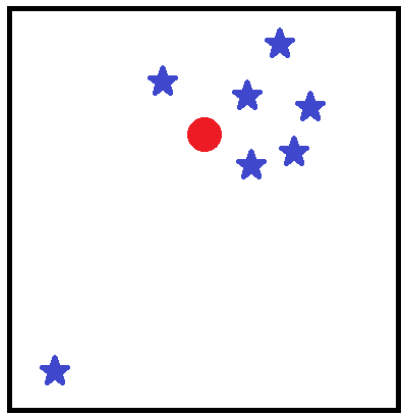
$$r_{\text{Bayes}}(\hat{y}) = \mathbb{E}_{y(x)|x}[\mathcal{L}(\hat{y}; y(x))]$$

$\hat{y}_{\text{Bayes}}(x)$ is a Bayes estimator if

$$\hat{y}_{\text{Bayes}}(x) = \operatorname*{argmin}_{\hat{y} \in \mathcal{Y}} r_{\text{Bayes}}(\hat{y})$$

Example: point estimation

- $\mathcal{Y} = \{\star\text{'s}\}$, $\hat{y}_{\text{minimax}} =$ ●
- Bayes estimator tries to accommodate everyone



Example: Alien attack!

Should I...

- stay and admire the beauty of nature
- run for my life

Example: Alien attack!

vs

- Bayes risk:

$$\text{risk} = \mathcal{L}(\text{death}) \times \underbrace{\Pr(\text{alien!})}_{=0.0001\%*} + \mathcal{L}(\text{boring night}) \times \underbrace{\Pr(\text{harmless meteor})}_{=99.999\%*}$$

- Minimax risk: There is a nonzero chance that the alien will eat me.

$$\text{risk} = \mathcal{L}(\text{death})$$

*I have no citation for this

Example: Alien attack!

vs

- Bayes risk:

$$\text{risk} = \mathcal{L}(\text{death}) \times \underbrace{\Pr(\text{alien!})}_{=0.0001\%*} + \mathcal{L}(\text{boring night}) \times \underbrace{\Pr(\text{harmless meteor})}_{=99.999\%*}$$

- Minimax risk: There is a nonzero chance that the alien will eat me.

$$\text{risk} = \mathcal{L}(\text{death})$$

- Bayes estimator:
- Minimax estimator:

*I have no citation for this

Example: Alien attack!

vs

- Bayes risk:

$$\text{risk} = \mathcal{L}(\text{death}) \times \underbrace{\Pr(\text{alien!})}_{=0.0001\%*} + \mathcal{L}(\text{boring night}) \times \underbrace{\Pr(\text{harmless meteor})}_{=99.999\%*}$$

- Minimax risk: There is a nonzero chance that the alien will eat me.

$$\text{risk} = \mathcal{L}(\text{death})$$

- Bayes estimator: stay
- Minimax estimator:

*I have no citation for this

Example: Alien attack!

vs

- Bayes risk:

$$\text{risk} = \mathcal{L}(\text{death}) \times \underbrace{\Pr(\text{alien!})}_{=0.0001\%*} + \mathcal{L}(\text{boring night}) \times \underbrace{\Pr(\text{harmless meteor})}_{=99.999\%*}$$

- Minimax risk: There is a nonzero chance that the alien will eat me.

$$\text{risk} = \mathcal{L}(\text{death})$$

- Bayes estimator: stay
- Minimax estimator: run!

*I have no citation for this

Bayes classifier in binary classification

Binary loss function: $\mathcal{L}(\hat{y}, y) = \begin{cases} 0 & \text{if } \hat{y} = y \\ 1 & \text{else.} \end{cases}$

- What is Bayes risk?

- What is Bayes classifier?

- What is Bayes risk of Bayes classifier??

Bayes classifier in binary classification

Binary loss function: $\mathcal{L}(\hat{y}, y) = \begin{cases} 0 & \text{if } \hat{y} = y \\ 1 & \text{else.} \end{cases}$

- What is Bayes risk?

$$r_{\text{Bayes}}(\hat{y}) := \mathbb{E}_{y(x)|x}[\mathcal{L}(y(x), \hat{y})] = \begin{cases} \Pr(y(x) = 1) & \text{if } \hat{y} = 0 \\ \Pr(y(x) = 0) & \text{if } \hat{y} = 1 \end{cases}$$

- What is Bayes classifier?

- What is Bayes risk of Bayes classifier??

Bayes classifier in binary classification

Binary loss function: $\mathcal{L}(\hat{y}, y) = \begin{cases} 0 & \text{if } \hat{y} = y \\ 1 & \text{else.} \end{cases}$

- What is Bayes risk?

$$r_{\text{Bayes}}(\hat{y}) := \mathbb{E}_{y(x)|x}[\mathcal{L}(y(x), \hat{y})] = \begin{cases} \Pr(y(x) = 1) & \text{if } \hat{y} = 0 \\ \Pr(y(x) = 0) & \text{if } \hat{y} = 1 \end{cases}$$

- What is Bayes classifier?

$$\hat{y}_{\text{Bayes}}(x) = \begin{cases} 1 & \text{if } \Pr(y(x) = 1) > 0.5 \\ 0 & \text{if } \Pr(y(x) = 1) < 0.5 \\ \text{either} & \text{if } \Pr(y(x) = 1) = 0.5 \end{cases}$$

- What is Bayes risk of Bayes classifier??

Bayes classifier in binary classification

Binary loss function: $\mathcal{L}(\hat{y}, y) = \begin{cases} 0 & \text{if } \hat{y} = y \\ 1 & \text{else.} \end{cases}$

- What is Bayes risk?

$$r_{\text{Bayes}}(\hat{y}) := \mathbb{E}_{y(x)|x}[\mathcal{L}(y(x), \hat{y})] = \begin{cases} \Pr(y(x) = 1) & \text{if } \hat{y} = 0 \\ \Pr(y(x) = 0) & \text{if } \hat{y} = 1 \end{cases}$$

- What is Bayes classifier?

$$\hat{y}_{\text{Bayes}}(x) = \begin{cases} 1 & \text{if } \Pr(y(x) = 1) > 0.5 \\ 0 & \text{if } \Pr(y(x) = 1) < 0.5 \\ \text{either} & \text{if } \Pr(y(x) = 1) = 0.5 \end{cases}$$

- What is Bayes risk of Bayes classifier??

$$r_{\text{Bayes}}(\hat{y}_{\text{Bayes}}) = \min(\Pr(y(x) = 1), 1 - \Pr(y(x) = 1))$$

Bayes classifier great, not practical

$$r_{\text{Bayes}}(\hat{y}(x)) = \mathbb{E}_{y(x)|x}[\mathcal{L}(\hat{y}(x); y(x))],$$

Optimality Bayes classifier minimizes expected loss

$$\begin{aligned}
\text{Expected loss} &= \mathbb{E}_{x,y(x)}[\mathcal{L}(\hat{y}_{\text{Bayes}}, y(x))] \\
&= \mathbb{E}_x[\underbrace{\mathbb{E}_{y(x)|x}[\mathcal{L}(\hat{y}_{\text{Bayes}}, y(x))]}_{\text{optimal}}] \\
&\leq \mathbb{E}_x[\mathbb{E}_{y(x)|x}[\mathcal{L}(\hat{y}, y(x))]], \text{ any scheme } \hat{y}(x)
\end{aligned}$$

Bayes classifier great, not practical

$$r_{\text{Bayes}}(\hat{y}(x)) = \mathbb{E}_{y(x)|x}[\mathcal{L}(\hat{y}(x); y(x))],$$

Optimality Bayes classifier minimizes expected loss

$$\begin{aligned}
\text{Expected loss} &= \mathbb{E}_{x,y(x)}[\mathcal{L}(\hat{y}_{\text{Bayes}}, y(x))] \\
&= \mathbb{E}_x[\underbrace{\mathbb{E}_{y(x)|x}[\mathcal{L}(\hat{y}_{\text{Bayes}}, y(x))]}_{\text{optimal}}] \\
&\leq \mathbb{E}_x[\mathbb{E}_{y(x)|x}[\mathcal{L}(\hat{y}, y(x))]], \text{ any scheme } \hat{y}(x)
\end{aligned}$$

Computational complexity

- Assume K classes, D attributes, each taking N values.
- Sampling complexity for $y|x$?
- Problem: no assumptions on $y|x$

Bayes classifier great, not practical

$$r_{\text{Bayes}}(\hat{y}(x)) = \mathbb{E}_{y(x)|x}[\mathcal{L}(\hat{y}(x); y(x))],$$

Optimality Bayes classifier minimizes expected loss

$$\begin{aligned}
\text{Expected loss} &= \mathbb{E}_{x,y(x)}[\mathcal{L}(\hat{y}_{\text{Bayes}}, y(x))] \\
&= \mathbb{E}_x[\underbrace{\mathbb{E}_{y(x)|x}[\mathcal{L}(\hat{y}_{\text{Bayes}}, y(x))]}_{\text{optimal}}] \\
&\leq \mathbb{E}_x[\mathbb{E}_{y(x)|x}[\mathcal{L}(\hat{y}, y(x))]], \text{ any scheme } \hat{y}(x)
\end{aligned}$$

Computational complexity

- Assume K classes, D attributes, each taking N values.
- Sampling complexity for $y|x$?
  Ans: $O(KN^D)$ (requires exponentially large training set)
- Problem: no assumptions on $y|x$

Nearest neighbor classifier

Nearest neighbor classifier

- All I get is some labeled samples
- No idea which features are more/less relevant
- Nearest neighbor (NN) classifier: label function $h(x)$

$$h(x) = h(x') : x' = \underset{x' \text{ labeled}}{\operatorname{argmin}} \|x - x'\|_2$$

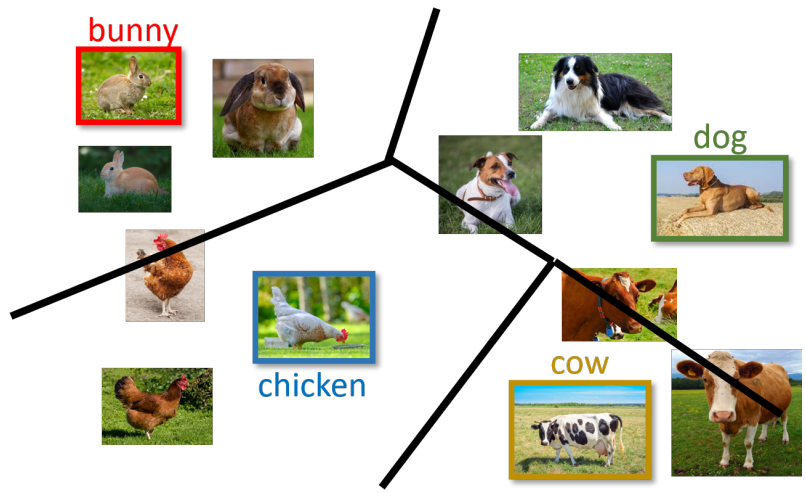
- Regions carved out by classifier are called Voronoi regions

Nearest neighbor classifier

- All I get is some labeled samples
- No idea which features are more/less relevant
- Nearest neighbor (NN) classifier: label function $h(x)$

$$h(x) = h(x') : x' = \underset{x' \text{ labeled}}{\operatorname{argmin}} \|x - x'\|_2$$

- Regions carved out by classifier are called Voronoi regions

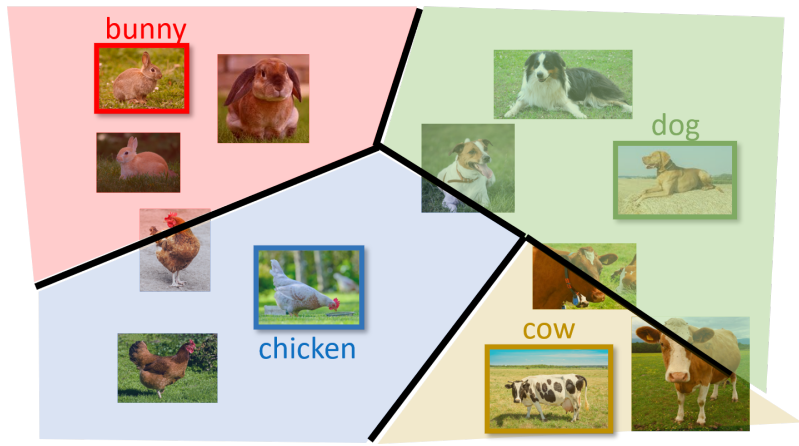


Nearest neighbor classifier

- All I get is some labeled samples
- No idea which features are more/less relevant
- Nearest neighbor (NN) classifier: label function $h(x)$

$$h(x) = h(x') : x' = \underset{x' \text{ labeled}}{\operatorname{argmin}} \|x - x'\|_2$$

- Regions carved out by classifier are called Voronoi regions



Nearest neighbor classifier

- Given $x_1, ..., x_m$ unlabeled, and $\tilde{x}_1, ..., \tilde{x}_{\tilde{m}}$ labeled, define

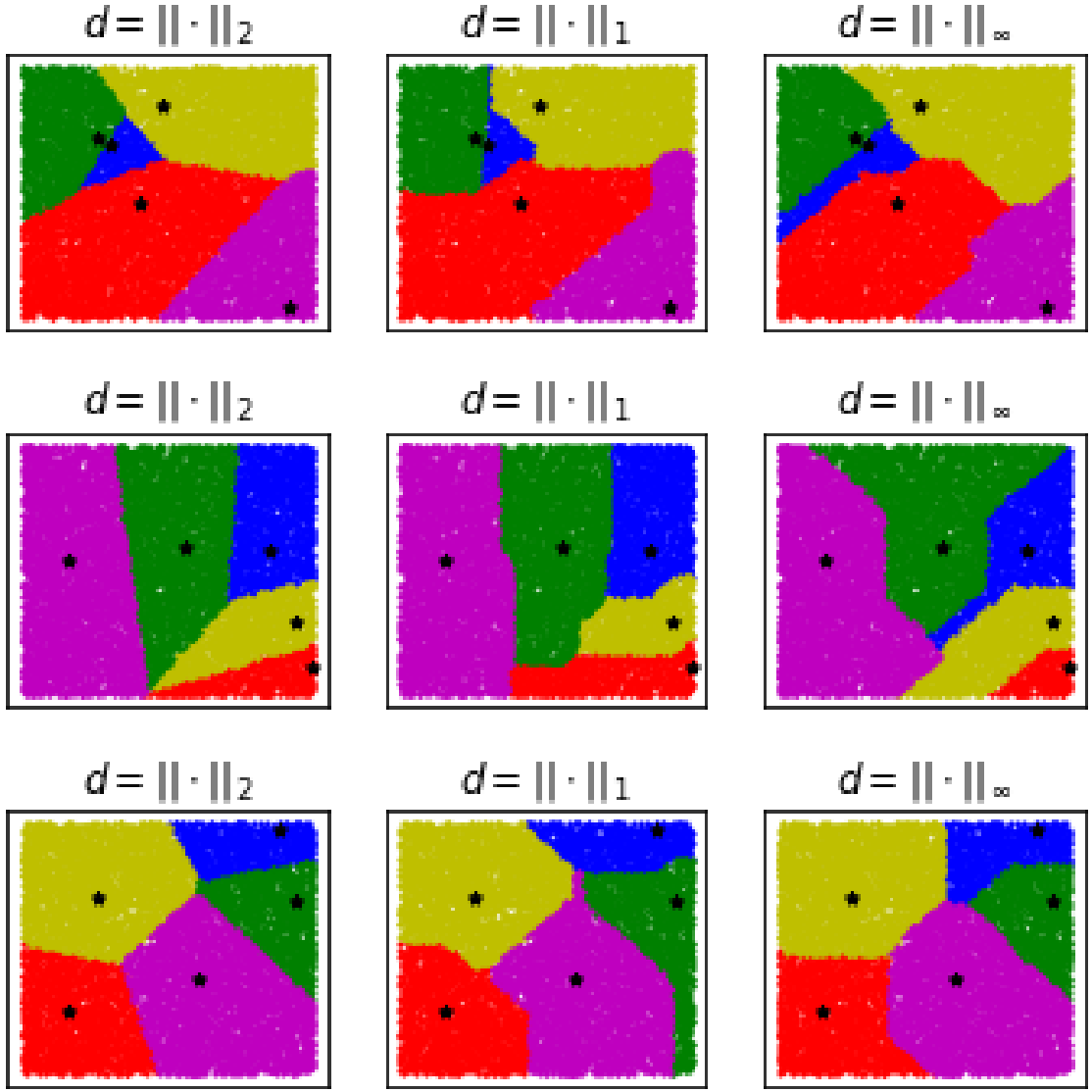
$$z_x := \underset{z,\ \text{labeled}}{\operatorname{argmin}} \{\|x - z\|\} \qquad \text{Nearest labeled point}$$

- The 1-nearest neighbor classifier picks the label of the closest labeled point

$$\hat{y}_{\mathsf{NN}}(x) := y(z_x),$$

- The K-nearest neighbor classifier aggregates labels of K closest points
  - Majority vote, or average value
- Note: different distance functions lead to different classifiers

1-NN with different distance functions

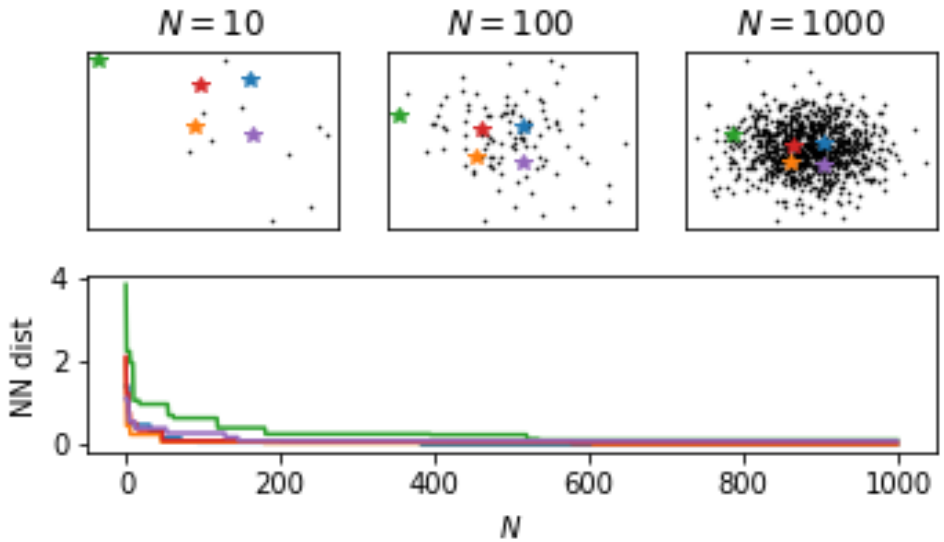


Asymptotic consistency of estimator

Claim:

- Suppose I sample points $x_i \sim \mathcal{D}$ for $i = 1, ..., N$, i.i.d.
- As $N \to +\infty$, every point approaches its neighbor.
- That is, pick any i, and for its closest neighbor,

$$\min_{\substack{j=1,...,N \\ j \neq i}} \|x_i - x_j\| \to 0$$



Asymptotic consistency of estimator

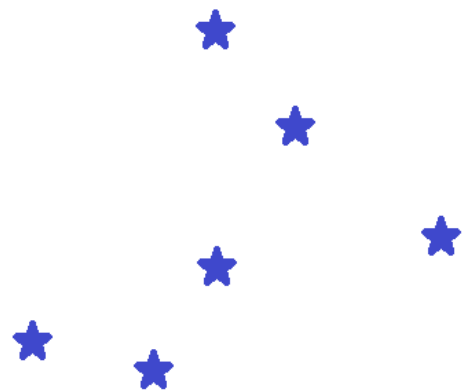
Claim: $\min_{\substack{j=1,\ldots,N \\ j\neq i}} \|x_i - x_j\| \overset{N\to+\infty}{\to} 0$

Proof:

- Define

$$R_\delta(x_0) = \{x : \|x_0 - x\|_2 \leq \delta\}, \qquad \epsilon := \Pr(\text{new } x \in R_\delta(x_0))$$

- Regularity assumption: $\epsilon > 0$



Asymptotic consistency of estimator

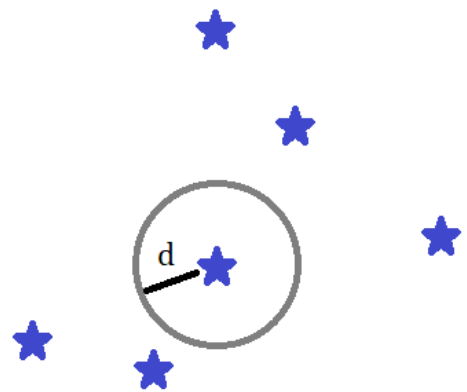
Claim: $\min_{\substack{j=1,\ldots,N \\ j\neq i}} \|x_i - x_j\| \overset{N\to+\infty}{\to} 0$

Proof:

- Define

$$R_\delta(x_0) = \{x : \|x_0 - x\|_2 \leq \delta\}, \qquad \epsilon := \Pr(\text{new } x \in R_\delta(x_0))$$

- Regularity assumption: $\epsilon > 0$



Asymptotic consistency of estimator

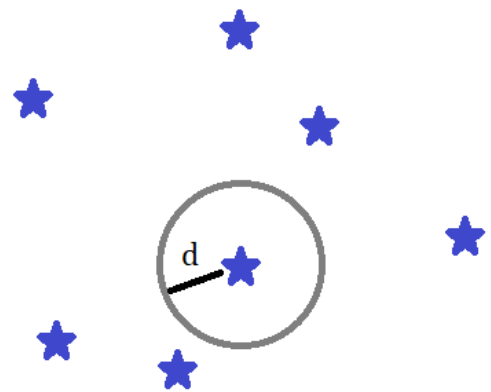
Claim: $\min_{\substack{j=1,\ldots,N \\ j\neq i}} \|x_i - x_j\| \overset{N\to+\infty}{\to} 0$

Proof:

- Define

$$R_\delta(x_0) = \{x : \|x_0 - x\|_2 \leq \delta\}, \qquad \epsilon := \Pr(\text{new } x \in R_\delta(x_0))$$

- Regularity assumption: $\epsilon > 0$



Asymptotic consistency of estimator

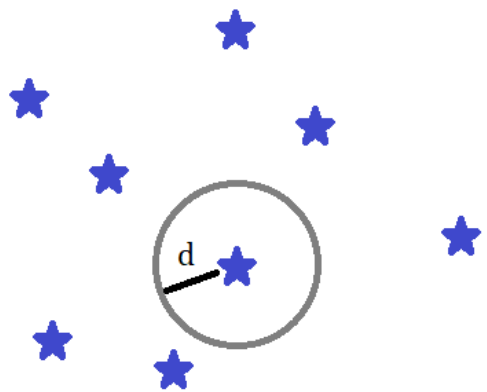
Claim: $\min_{\substack{j=1,\ldots,N \\ j\neq i}} \|x_i - x_j\| \overset{N\to+\infty}{\to} 0$

Proof:

- Define

$$R_\delta(x_0) = \{x : \|x_0 - x\|_2 \leq \delta\}, \qquad \epsilon := \Pr(\text{new } x \in R_\delta(x_0))$$

- Regularity assumption: $\epsilon > 0$



Asymptotic consistency of estimator

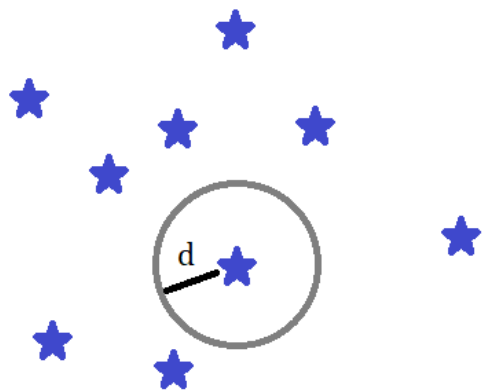
Claim: $\min_{\substack{j=1,\ldots,N \\ j\neq i}} \|x_i - x_j\| \overset{N\to+\infty}{\to} 0$

Proof:

- Define

$$R_\delta(x_0) = \{x : \|x_0 - x\|_2 \leq \delta\}, \qquad \epsilon := \Pr(\text{new } x \in R_\delta(x_0))$$

- Regularity assumption: $\epsilon > 0$



Asymptotic consistency of estimator

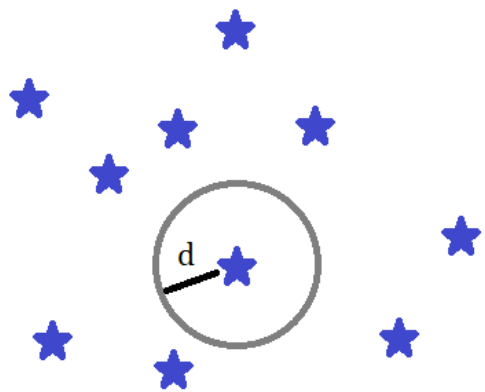
Claim: $\min_{\substack{j=1,\ldots,N \\ j\neq i}} \|x_i - x_j\| \overset{N\to+\infty}{\to} 0$

Proof:

- Define

$$R_\delta(x_0) = \{x : \|x_0 - x\|_2 \leq \delta\}, \qquad \epsilon := \Pr(\text{new } x \in R_\delta(x_0))$$

- Regularity assumption: $\epsilon > 0$



Asymptotic consistency of estimator

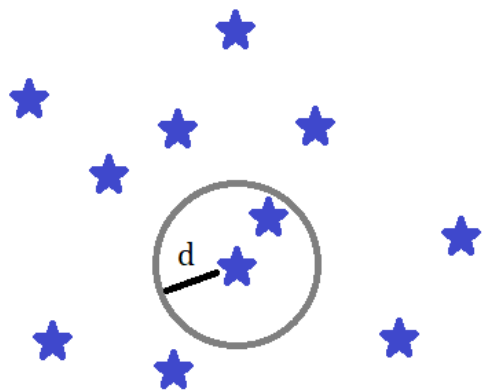
Claim: $\min_{\substack{j=1,\ldots,N \\ j\neq i}} \|x_i - x_j\| \overset{N\to+\infty}{\to} 0$

Proof:

- Define

$$R_\delta(x_0) = \{x : \|x_0 - x\|_2 \leq \delta\}, \qquad \epsilon := \Pr(\text{new } x \in R_\delta(x_0))$$

- Regularity assumption: $\epsilon > 0$



Asymptotic consistency of estimator

Claim: $\min_{\substack{j=1,\ldots,N \\ j \neq i}} \|x_i - x_j\| \to 0$

Proof:

- What are the chances that next $x_i \notin R_\delta(x_0)$?

Asymptotic consistency of estimator

Claim: $\min_{\substack{j=1,\ldots,N \\ j \neq i}} \|x_i - x_j\| \to 0$

Proof:

- What are the chances that next $x_i \notin R_\delta(x_0)$?

$$\Pr(x_i \notin R_\delta(x_0)) = 1 - \Pr(x_i \in R_\delta(x_0)) = \epsilon \quad \text{(by definition of } \epsilon\text{)}$$

Asymptotic consistency of estimator

Claim: $\min_{\substack{j=1,...,N \\ j \neq i}} \|x_i - x_j\| \to 0$

Proof:

- What are the chances that next $x_i \notin R_\delta(x_0)$?

$$\Pr(x_i \notin R_\delta(x_0)) = 1 - \Pr(x_i \in R_\delta(x_0)) = \epsilon \quad \text{(by definition of } \epsilon\text{)}$$

- What are the chances that no x_i, for all $i = 1, ..., N$, are in $R_\delta(x_0)$?

Asymptotic consistency of estimator

Claim: $\min_{\substack{j=1,...,N \\ j \neq i}} \|x_i - x_j\| \to 0$

Proof:

- What are the chances that next $x_i \notin R_\delta(x_0)$?

$$\Pr(x_i \notin R_\delta(x_0)) = 1 - \Pr(x_i \in R_\delta(x_0)) = \epsilon \quad \text{(by definition of } \epsilon\text{)}$$

- What are the chances that no x_i, for all $i = 1, ..., N$, are in $R_\delta(x_0)$?

$$\Pr(x_i \notin R_\delta(x_0),\ \forall i = 1, ..., N) = (1 - \Pr(x_i \in R_\delta(x_0)))^N$$

Asymptotic consistency of estimator

Claim: $\min_{\substack{j=1,...,N \\ j \neq i}} \|x_i - x_j\| \to 0$

Proof:

- What are the chances that next $x_i \notin R_\delta(x_0)$?

$$\Pr(x_i \notin R_\delta(x_0)) = 1 - \Pr(x_i \in R_\delta(x_0)) = \epsilon \quad \text{(by definition of } \epsilon\text{)}$$

- What are the chances that no x_i, for all $i = 1, ..., N$, are in $R_\delta(x_0)$?

$$\Pr(x_i \notin R_\delta(x_0),\ \forall i = 1, ..., N) = (1 - \Pr(x_i \in R_\delta(x_0)))^N$$

- As $N \to \infty$, what are the chances that there are no points $x_1, ..., x_N$ whose distance to x_0 is $< \delta$?

Asymptotic consistency of estimator

Claim: $\min_{\substack{j=1,...,N \\ j \neq i}} \|x_i - x_j\| \to 0$

Proof:

- What are the chances that next $x_i \notin R_\delta(x_0)$?

$$\Pr(x_i \notin R_\delta(x_0)) = 1 - \Pr(x_i \in R_\delta(x_0)) = \epsilon \quad \text{(by definition of } \epsilon\text{)}$$

- What are the chances that no x_i, for all $i = 1, ..., N$, are in $R_\delta(x_0)$?

$$\Pr(x_i \notin R_\delta(x_0),\ \forall i = 1, ..., N) = (1 - \Pr(x_i \in R_\delta(x_0)))^N$$

- As $N \to \infty$, what are the chances that there are no points $x_1, ..., x_N$ whose distance to x_0 is $< \delta$?

$$(1 - \Pr(x_i \in R_\delta(x_0)))^N \overset{N \to \infty}{\to} 0$$

Asymptotic consistency of estimator

Claim: $\min_{\substack{j=1,...,N \\ j \neq i}} \|x_i - x_j\| \to 0$

Proof:

- What are the chances that next $x_i \notin R_\delta(x_0)$?

$$\Pr(x_i \notin R_\delta(x_0)) = 1 - \Pr(x_i \in R_\delta(x_0)) = \epsilon \quad \text{(by definition of } \epsilon\text{)}$$

- What are the chances that no x_i, for all $i = 1, ..., N$, are in $R_\delta(x_0)$?

$$\Pr(x_i \notin R_\delta(x_0),\ \forall i = 1, ..., N) = (1 - \Pr(x_i \in R_\delta(x_0)))^N$$

- As $N \to \infty$, what are the chances that there are no points $x_1, ..., x_N$ whose distance to x_0 is $< \delta$?

$$(1 - \Pr(x_i \in R_\delta(x_0)))^N \overset{N \to \infty}{\to} 0$$

- This holds true for arbitrarily small $\delta > 0$

Bayes Risk for 1-nearest neighbor

Define z_x the labeled point closest to x

$$
\begin{aligned}
r_{\text{Bayes}}(\hat{y}_{\text{NN}}(x)) &= \mathbb{E}_{y(x)|x}[\mathcal{L}(\underbrace{\hat{y}_{\text{NN}}(x)}_{=y(z_x)}, y(x))] \\
&= \Pr(y(x)=1, y(z_x)=0|x) \\
&\quad + \Pr(y(x)=0, y(z_x)=1|x) \\
&\overset{\text{i.i.d.}}{=} \Pr(y(x)=1|x)\Pr(y(z_x)=0|x) \\
&\quad + \Pr(y(x)=0|x)\Pr(y(z_x)=1|x)
\end{aligned}
$$

As $N \to +\infty$, $z_x \to x$. Then $\Pr(y(z_x)) \to \Pr(y(x))$.

$$
\lim_{N \to +\infty} r_{\text{Bayes}}(\hat{y}_{\text{NN}}(x)) = 2 \cdot (\Pr(y(x)=1|x)) \cdot (1 - \Pr(y(x)=1|x))
$$

Bounds on 1-NN Bayes risk at asymptotic limit

- Define $\eta(x) := \Pr(y(x) = 1|x)$.
- At asymptotic limit, for a single point,

$$r_{\text{Bayes}}(\hat{y}_{\text{Bayes}}) = \min\{\eta(x), 1 - \eta(x)\}, \quad r_{\text{Bayes}}(\hat{y}_{\text{NN}}(x)) = 2\eta(x)(1 - \eta(x))$$

- Over all points, define

$$R^* = \mathbb{E}_x[r_{\text{Bayes}}(\hat{y}_{\text{Bayes}})], \quad R_{\text{NN}} = \mathbb{E}_x[r_{\text{Bayes}}(\hat{y}_{\text{NN}}(x))]$$

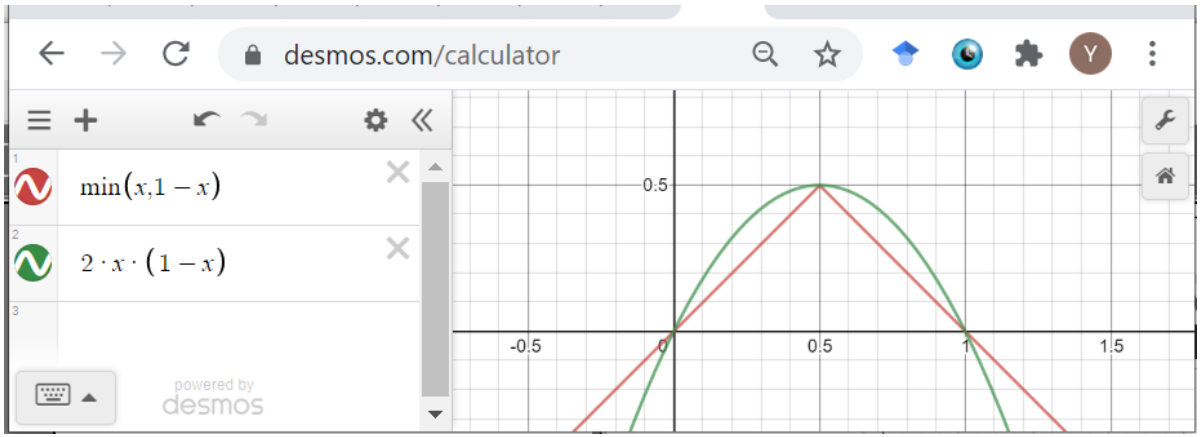
- Bounds:

$$R^* \leq R_{\text{NN}} \leq 2R^*(1 - R^*)$$

Bounds derivation

- $\eta(x) := \Pr(y(x) = 1|x)$, $\beta(x) = \min\{\eta(x), 1 - \eta(x)\}$.
- $r_{\text{Bayes}}(\hat{y}_{\text{Bayes}}) = \beta(x)$, $r_{\text{Bayes}}(\hat{y}_{\text{NN}}(x)) = 2\beta(x)(1 - \beta(x))$
- $R^* = \mathbb{E}_x[r_{\text{Bayes}}(\hat{y}_{\text{Bayes}})]$, $R_{\text{NN}} = \mathbb{E}_x[r_{\text{Bayes}}(\hat{y}_{\text{NN}}(x))]$

Proof of lower bound For any $0 \leq \eta \leq 1$, $\min\{\eta, 1 - \eta\} \leq 2\eta(1 - \eta)$



This shows $R^* \leq R_{\text{NN}}$

Bounds derivation

- $\eta(x) := \Pr(y(x) = 1|x)$, $\beta(x) = \min\{\eta(x), 1 - \eta(x)\}$.
- $r_{\mathsf{Bayes}}(\hat{y}_{\mathsf{Bayes}}) = \beta(x)$, $r_{\mathsf{Bayes}}(\hat{y}_{\mathsf{NN}}(x)) = 2\beta(x)(1 - \beta(x))$
- $R^* = \mathbb{E}_x[r_{\mathsf{Bayes}}(\hat{y}_{\mathsf{Bayes}})]$, $R_{\mathsf{NN}} = \mathbb{E}_x[r_{\mathsf{Bayes}}(\hat{y}_{\mathsf{NN}}(x))]$

Proof of upper bound Recall definition of variance

$$0 \leq \mathbf{var}_x[\beta]$$

Bounds derivation

- $\eta(x) := \Pr(y(x) = 1|x)$, $\beta(x) = \min\{\eta(x), 1 - \eta(x)\}$.
- $r_{\mathsf{Bayes}}(\hat{y}_{\mathsf{Bayes}}) = \beta(x)$, $r_{\mathsf{Bayes}}(\hat{y}_{\mathsf{NN}}(x)) = 2\beta(x)(1 - \beta(x))$
- $R^* = \mathbb{E}_x[r_{\mathsf{Bayes}}(\hat{y}_{\mathsf{Bayes}})]$, $R_{\mathsf{NN}} = \mathbb{E}_x[r_{\mathsf{Bayes}}(\hat{y}_{\mathsf{NN}}(x))]$

Proof of upper bound Recall definition of variance

$$0 \leq \mathbf{var}_x[\beta] = \mathbb{E}_x[(\beta - \mathbb{E}_x[\beta])^2]$$

Bounds derivation

- $\eta(x) := \Pr(y(x) = 1|x)$, $\beta(x) = \min\{\eta(x), 1 - \eta(x)\}$.
- $r_{\mathsf{Bayes}}(\hat{y}_{\mathsf{Bayes}}) = \beta(x)$, $r_{\mathsf{Bayes}}(\hat{y}_{\mathsf{NN}}(x)) = 2\beta(x)(1 - \beta(x))$
- $R^* = \mathbb{E}_x[r_{\mathsf{Bayes}}(\hat{y}_{\mathsf{Bayes}})]$, $R_{\mathsf{NN}} = \mathbb{E}_x[r_{\mathsf{Bayes}}(\hat{y}_{\mathsf{NN}}(x))]$

Proof of upper bound Recall definition of variance

$$0 \leq \mathbf{var}_x[\beta] = \mathbb{E}_x[(\beta - \mathbb{E}_x[\beta])^2] = \mathbb{E}_x[\beta^2] - (\mathbb{E}_x[\beta])^2$$

Bounds derivation

- $\eta(x) := \Pr(y(x) = 1|x)$, $\beta(x) = \min\{\eta(x), 1 - \eta(x)\}$.
- $r_{\text{Bayes}}(\hat{y}_{\text{Bayes}}) = \beta(x)$, $r_{\text{Bayes}}(\hat{y}_{\text{NN}}(x)) = 2\beta(x)(1 - \beta(x))$
- $R^* = \mathbb{E}_x[r_{\text{Bayes}}(\hat{y}_{\text{Bayes}})]$, $R_{\text{NN}} = \mathbb{E}_x[r_{\text{Bayes}}(\hat{y}_{\text{NN}}(x))]$

Proof of upper bound Recall definition of variance

$$0 \leq \mathbf{var}_x[\beta] = \mathbb{E}_x[(\beta - \mathbb{E}_x[\beta])^2] = \mathbb{E}_x[\beta^2] - (\mathbb{E}_x[\beta])^2$$

which shows that $\mathbb{E}[\beta^2] \geq (\mathbb{E}[\beta])^2$.

Bounds derivation

- $\eta(x) := \Pr(y(x) = 1|x)$, $\beta(x) = \min\{\eta(x), 1 - \eta(x)\}$.
- $r_{\text{Bayes}}(\hat{y}_{\text{Bayes}}) = \beta(x)$, $r_{\text{Bayes}}(\hat{y}_{\text{NN}}(x)) = 2\beta(x)(1 - \beta(x))$
- $R^* = \mathbb{E}_x[r_{\text{Bayes}}(\hat{y}_{\text{Bayes}})]$, $R_{\text{NN}} = \mathbb{E}_x[r_{\text{Bayes}}(\hat{y}_{\text{NN}}(x))]$

Proof of upper bound Recall definition of variance

$$0 \leq \mathbf{var}_x[\beta] = \mathbb{E}_x[(\beta - \mathbb{E}_x[\beta])^2] = \mathbb{E}_x[\beta^2] - (\mathbb{E}_x[\beta])^2$$

which shows that $\mathbb{E}[\beta^2] \geq (\mathbb{E}[\beta])^2$. Then

$$R_{\text{NN}} \quad = \quad \mathbb{E}[2\beta(1 - \beta)] = 2\mathbb{E}[\beta] - 2\mathbb{E}[\beta^2]$$

Bounds derivation

- $\eta(x) := \Pr(y(x) = 1|x)$, $\beta(x) = \min\{\eta(x), 1 - \eta(x)\}$.
- $r_{\mathsf{Bayes}}(\hat{y}_{\mathsf{Bayes}}) = \beta(x)$, $r_{\mathsf{Bayes}}(\hat{y}_{\mathsf{NN}}(x)) = 2\beta(x)(1 - \beta(x))$
- $R^* = \mathbb{E}_x[r_{\mathsf{Bayes}}(\hat{y}_{\mathsf{Bayes}})]$, $R_{\mathsf{NN}} = \mathbb{E}_x[r_{\mathsf{Bayes}}(\hat{y}_{\mathsf{NN}}(x))]$

Proof of upper bound Recall definition of variance

$$0 \leq \mathbf{var}_x[\beta] = \mathbb{E}_x[(\beta - \mathbb{E}_x[\beta])^2] = \mathbb{E}_x[\beta^2] - (\mathbb{E}_x[\beta])^2$$

which shows that $\mathbb{E}[\beta^2] \geq (\mathbb{E}[\beta])^2$. Then

$$\begin{aligned} R_{\mathsf{NN}} &= \mathbb{E}[2\beta(1 - \beta)] = 2\mathbb{E}[\beta] - 2\mathbb{E}[\beta^2] \\ &\leq 2\mathbb{E}[\beta] - 2(\mathbb{E}[\beta])^2 \end{aligned}$$

Bounds derivation

- $\eta(x) := \Pr(y(x) = 1|x)$, $\beta(x) = \min\{\eta(x), 1 - \eta(x)\}$.
- $r_{\mathsf{Bayes}}(\hat{y}_{\mathsf{Bayes}}) = \beta(x)$, $r_{\mathsf{Bayes}}(\hat{y}_{\mathsf{NN}}(x)) = 2\beta(x)(1 - \beta(x))$
- $R^* = \mathbb{E}_x[r_{\mathsf{Bayes}}(\hat{y}_{\mathsf{Bayes}})]$, $R_{\mathsf{NN}} = \mathbb{E}_x[r_{\mathsf{Bayes}}(\hat{y}_{\mathsf{NN}}(x))]$

Proof of upper bound Recall definition of variance

$$0 \leq \mathbf{var}_x[\beta] = \mathbb{E}_x[(\beta - \mathbb{E}_x[\beta])^2] = \mathbb{E}_x[\beta^2] - (\mathbb{E}_x[\beta])^2$$

which shows that $\mathbb{E}[\beta^2] \geq (\mathbb{E}[\beta])^2$. Then

$$\begin{aligned} R_{\mathsf{NN}} &= \mathbb{E}[2\beta(1 - \beta)] = 2\mathbb{E}[\beta] - 2\mathbb{E}[\beta^2] \\ &\leq 2\mathbb{E}[\beta] - 2(\mathbb{E}[\beta])^2 \\ &= 2R^*(1 - R^*) \end{aligned}$$

Bounds derivation

- $\eta(x) := \Pr(y(x) = 1|x)$, $\beta(x) = \min\{\eta(x), 1 - \eta(x)\}$.
- $r_{\text{Bayes}}(\hat{y}_{\text{Bayes}}) = \beta(x)$, $r_{\text{Bayes}}(\hat{y}_{\text{NN}}(x)) = 2\beta(x)(1 - \beta(x))$
- $R^* = \mathbb{E}_x[r_{\text{Bayes}}(\hat{y}_{\text{Bayes}})]$, $R_{\text{NN}} = \mathbb{E}_x[r_{\text{Bayes}}(\hat{y}_{\text{NN}}(x))]$

These bounds are tight

$$R^* \leq R_{\text{NN}} \leq 2R^*(1 - R^*)$$

- Suppose $R^* = 0$ (perfect recovery).
  - Then $0 = R^* = 2R^*(1 - R^*)$ and $R_{\text{NN}} = 0$.
- Suppose $R^* = \frac{1}{2}$ (no correlation between label and data)
  - Then $\frac{1}{2} = R^* = 2R^*(1 - R^*)$and $R_{\text{NN}} = \frac{1}{2}$.

Summary (1 of 3): Decision theory and Bayes risk

- Statistical decision theory: make choices to minimize loss
    - Definitions apply to any loss function, to many applications
- (Mini)max risk, Bayes risk
- Minimax estimator, Bayes estimator
    - Not practical, more analytical
- Computed Bayes risk of Bayes estimator

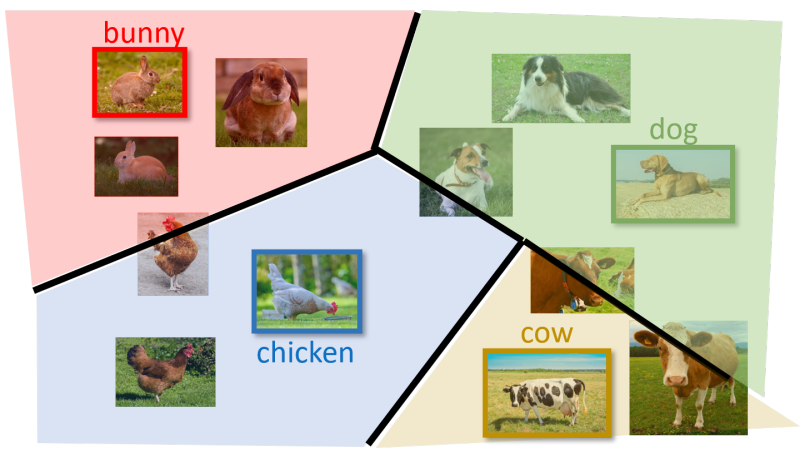
$$r_{\text{Bayes}}(\hat{y}_{\text{Bayes}}) = \beta(x)(1-\beta(x)), \quad \beta(x) := \min\{\Pr(y=1|x), \Pr(y=0|x)\}$$

Summary (2 of 3): K-nearest neighbors

- 1-NN

$$\hat{y}_{\text{NN}}(x) = y(z_x), \quad z_x := \text{closest labeled point to } x$$

- K-NN
  - Use averaging or majority vote to decide amongst K closest labeled points
- Different distance metrics lead to different classifiers



Summary (3 of 3): K-nearest neighbors

Asymptotic consistency: what happens when # data samples $N \to +\infty$?

- Nearest neighbors $z_x \to x$
- Bayes risk of nearest 1-neighbor over 1 point

$$r_{\text{Bayes}}(\hat{y}_{\text{NN}}(x)) = 2\eta(x)(1-\eta(x))$$

- Bayes risk of 1-nearest neighbor over distribution of x

$$R^* \leq R_{\text{NN}} \leq 2R^*(1-R^*)$$

  where

$$R^* = \mathbb{E}_x[r_{\text{Bayes}}(\hat{y}_{\text{Bayes}}(x))], \quad R_{\text{NN}} = \mathbb{E}_x[r_{\text{Bayes}}(\hat{y}_{\text{NN}}(x))]$$

- These bounds are tight