Outline

Formulation

Formulation

Given a query $q = (\mathbf{x}_q, \omega_q)$, the NN rule assigns the label of the closest sample in the training set $\mathcal{T}$:

Formulation

Given a query $q = (\mathbf{x}_q, \omega_q)$, the NN rule assigns the label of the closest sample in the training set $\mathcal{T}$:

$$\hat{\omega}_q = \omega_i : \underset{1 \leq i \leq |\mathcal{T}|}{\arg\min}\, D(\mathbf{x}_q, \mathbf{x}_i)$$

where $D : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$

Formulation

Given a query $q = (\mathbf{x}_q, \omega_q)$, the NN rule assigns the label of the closest sample in the training set $\mathcal{T}$:

$$\hat{\omega}_q = \omega_i : \underset{1 \leq i \leq |\mathcal{T}|}{\arg\min}\, D(\mathbf{x}_q, \mathbf{x}_i)$$

where $D : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$

Main features:

- Only requires the definition of a (dis)similarity measure:
  - ✓ Useful for feature-based and structural representations

Formulation

Given a query $q = (\mathbf{x}_q, \omega_q)$, the NN rule assigns the label of the closest sample in the training set $\mathcal{T}$:

$$\hat{\omega}_q = \omega_i : \underset{1 \leq i \leq |\mathcal{T}|}{\arg\min}\, D(\mathbf{x}_q, \mathbf{x}_i)$$

where $D : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$

Main features:

- Only requires the definition of a (dis)similarity measure:
  - ✓ Useful for feature-based and structural representations
- Does not derive a model out of the $\mathcal{T}$:
  - ✓ Adaptive
  - ✗ Inefficient as $|\mathcal{T}| \uparrow\uparrow$

Voronoi tessellation

Division of the feature space into regions

- Any point in that falls in a region is closer to the sample that defines it than to any other point/region

Voronoi tessellation

Division of the feature space into regions

- Any point in that falls in a region is closer to the sample that defines it than to any other point/region

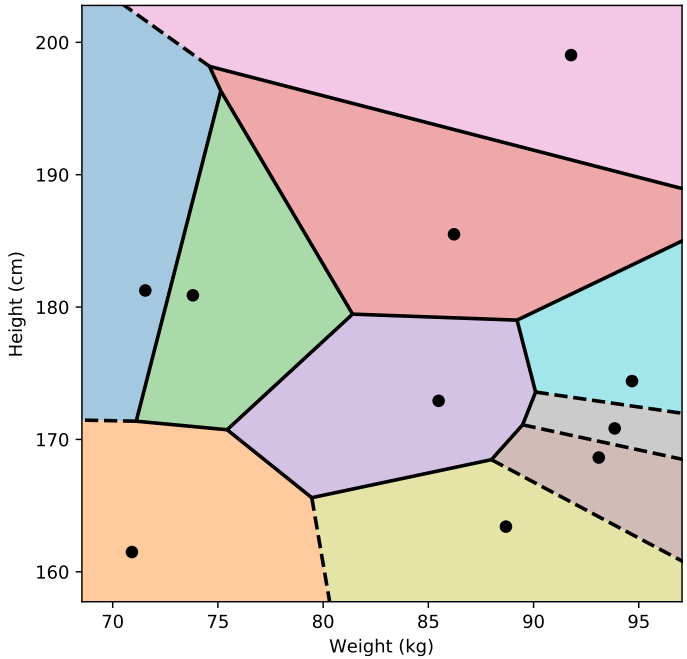
$$V_i = \left\{ \mathbf{x} \in \mathbb{R}^d : D(\mathbf{x}, \mathbf{x}_i) \leq D(\mathbf{x}, \mathbf{x}_j), \forall j \neq i \right\}$$

Voronoi tessellation

Division of the feature space into regions

- Any point in that falls in a region is closer to the sample that defines it than to any other point/region

$$V_i = \left\{ \mathbf{x} \in \mathbb{R}^d : D(\mathbf{x}, \mathbf{x}_i) \leq D(\mathbf{x}, \mathbf{x}_j), \forall j \neq i \right\}$$



Error rate and bounds

Assume a binary classification case $\Rightarrow \mathcal{W} = \{\omega_1, \omega_2\}$

Error rate and bounds

Assume a binary classification case $\Rightarrow \mathcal{W} = \{\omega_1, \omega_2\}$

- Bayes risk at $\mathbf{x}$: $R^*(\mathbf{x}) = \min \left[P(\omega_1|\mathbf{x}), \underbrace{P(\omega_2|\mathbf{x})}_{1-P(\omega_1|\mathbf{x})} \right]$
- Bayes risk: $\int_{-\infty}^{\infty} R^*(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$

Error rate and bounds

Assume a binary classification case $\Rightarrow \mathcal{W} = \{\omega_1, \omega_2\}$

- Bayes risk at $\mathbf{x}$: $R^*(\mathbf{x}) = \min \left[P(\omega_1|\mathbf{x}), \underbrace{P(\omega_2|\mathbf{x})}_{1-P(\omega_1|\mathbf{x})} \right]$
- Bayes risk: $\int_{-\infty}^{\infty} R^*(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$

Which is the probability of error at $\mathbf{x}$ for NN?

Error rate and bounds

Assume a binary classification case $\Rightarrow \mathcal{W} = \{\omega_1, \omega_2\}$

- Bayes risk at $\mathbf{x}$: $R^*(\mathbf{x}) = \min \left[P(\omega_1|\mathbf{x}), \underbrace{P(\omega_2|\mathbf{x})}_{1-P(\omega_1|\mathbf{x})} \right]$
- Bayes risk: $\int_{-\infty}^{\infty} R^*(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$

Which is the probability of error at $\mathbf{x}$ for NN?

$\rightarrow$ Assume $\mathbf{x}_N$ is the closest neighbor to $\mathbf{x}$

Error rate and bounds

Assume a binary classification case $\Rightarrow \mathcal{W} = \{\omega_1, \omega_2\}$

- Bayes risk at $\mathbf{x}$: $R^*(\mathbf{x}) = \min \left[P(\omega_1|\mathbf{x}), \underbrace{P(\omega_2|\mathbf{x})}_{1-P(\omega_1|\mathbf{x})} \right]$
- Bayes risk: $\int_{-\infty}^{\infty} R^*(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$

Which is the probability of error at $\mathbf{x}$ for NN?

$\rightarrow$ Assume $\mathbf{x}_N$ is the closest neighbor to $\mathbf{x}$

$R_{\text{NN}}(\mathbf{x}) = P(\omega \neq \hat{\omega}|\mathbf{x})$

Error rate and bounds

Assume a binary classification case $\Rightarrow \mathcal{W} = \{\omega_1, \omega_2\}$

- Bayes risk at $\mathbf{x}$: $R^*(\mathbf{x}) = \min \left[P(\omega_1|\mathbf{x}), \underbrace{P(\omega_2|\mathbf{x})}_{1-P(\omega_1|\mathbf{x})} \right]$
- Bayes risk: $\int_{-\infty}^{\infty} R^*(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$

Which is the probability of error at $\mathbf{x}$ for NN?

$\rightarrow$ Assume $\mathbf{x}_N$ is the closest neighbor to $\mathbf{x}$

$$R_{\text{NN}}(\mathbf{x}) = P(\omega \neq \hat{\omega}|\mathbf{x}) = P(\omega = \omega_1, \hat{\omega} = \omega_2|\mathbf{x}) + P(\omega = \omega_2, \hat{\omega} = \omega_1|\mathbf{x})$$

Error rate and bounds

Assume a binary classification case $\Rightarrow \mathcal{W} = \{\omega_1, \omega_2\}$

- Bayes risk at $\mathbf{x}$: $R^*(\mathbf{x}) = \min \left[P(\omega_1|\mathbf{x}), \underbrace{P(\omega_2|\mathbf{x})}_{1-P(\omega_1|\mathbf{x})} \right]$
- Bayes risk: $\int_{-\infty}^{\infty} R^*(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$

Which is the probability of error at $\mathbf{x}$ for NN?

$\rightarrow$ Assume $\mathbf{x}_N$ is the closest neighbor to $\mathbf{x}$

$$R_{\text{NN}}(\mathbf{x}) = P(\omega \neq \hat{\omega}|\mathbf{x}) = P(\omega = \omega_1, \hat{\omega} = \omega_2|\mathbf{x}) + P(\omega = \omega_2, \hat{\omega} = \omega_1|\mathbf{x}) =$$
$$= P(\omega_1|\mathbf{x}) \cdot P(\omega_2|\mathbf{x}_N) + P(\omega_2|\mathbf{x}) \cdot P(\omega_1|\mathbf{x}_N)$$

Error rate and bounds

If $|\mathcal{T}| \to \infty \Rightarrow P(\omega_i|\mathbf{x}_N) \approx P(\omega_i|\mathbf{x})$

Error rate and bounds

If $|\mathcal{T}| \to \infty \Rightarrow P(\omega_i|\mathbf{x}_N) \approx P(\omega_i|\mathbf{x})$

$$R_{\text{NN}}(\mathbf{x}) = P(\omega_1|\mathbf{x}) \cdot P(\omega_2|\mathbf{x}_N) + P(\omega_2|\mathbf{x}) \cdot P(\omega_1|\mathbf{x}_N) \approx$$
$$= P(\omega_1|\mathbf{x}) \cdot P(\omega_2|\mathbf{x}) + P(\omega_2|\mathbf{x}) \cdot P(\omega_1|\mathbf{x})$$

Error rate and bounds

If $|\mathcal{T}| \to \infty \Rightarrow P(\omega_i|\mathbf{x}_N) \approx P(\omega_i|\mathbf{x})$

$$
\begin{aligned}
R_{\text{NN}}(\mathbf{x}) &= P(\omega_1|\mathbf{x}) \cdot P(\omega_2|\mathbf{x}_N) + P(\omega_2|\mathbf{x}) \cdot P(\omega_1|\mathbf{x}_N) \approx \\
&= P(\omega_1|\mathbf{x}) \cdot P(\omega_2|\mathbf{x}) + P(\omega_2|\mathbf{x}) \cdot P(\omega_1|\mathbf{x}) = \\
&= 2 \cdot P(\omega_1|\mathbf{x}) \cdot P(\omega_2|\mathbf{x}) = 2 \cdot P(\omega_1|\mathbf{x}) \left[1 - P(\omega_1|\mathbf{x})\right]
\end{aligned}
$$

Error rate and bounds

If $|\mathcal{T}| \to \infty \Rightarrow P(\omega_i|\mathbf{x}_N) \approx P(\omega_i|\mathbf{x})$

$$
\begin{aligned}
R_{\text{NN}}(\mathbf{x}) &= P(\omega_1|\mathbf{x}) \cdot P(\omega_2|\mathbf{x}_N) + P(\omega_2|\mathbf{x}) \cdot P(\omega_1|\mathbf{x}_N) \approx \\
&= P(\omega_1|\mathbf{x}) \cdot P(\omega_2|\mathbf{x}) + P(\omega_2|\mathbf{x}) \cdot P(\omega_1|\mathbf{x}) = \\
&= 2 \cdot P(\omega_1|\mathbf{x}) \cdot P(\omega_2|\mathbf{x}) = 2 \cdot P(\omega_1|\mathbf{x}) \left[1 - P(\omega_1|\mathbf{x})\right] = \\
&= 2 \cdot R^*(\mathbf{x}) \left[1 - R^*(\mathbf{x})\right]
\end{aligned}
$$

Error rate and bounds

If $|\mathcal{T}| \to \infty \Rightarrow P(\omega_i|\mathbf{x}_N) \approx P(\omega_i|\mathbf{x})$

$$\begin{aligned} R_{\text{NN}}(\mathbf{x}) &= P(\omega_1|\mathbf{x}) \cdot P(\omega_2|\mathbf{x}_N) + P(\omega_2|\mathbf{x}) \cdot P(\omega_1|\mathbf{x}_N) \approx \\ &= P(\omega_1|\mathbf{x}) \cdot P(\omega_2|\mathbf{x}) + P(\omega_2|\mathbf{x}) \cdot P(\omega_1|\mathbf{x}) = \\ &= 2 \cdot P(\omega_1|\mathbf{x}) \cdot P(\omega_2|\mathbf{x}) = 2 \cdot P(\omega_1|\mathbf{x}) \left[1 - P(\omega_1|\mathbf{x})\right] = \\ &= 2 \cdot R^*(\mathbf{x}) \left[1 - R^*(\mathbf{x})\right] \end{aligned}$$

Error bound for NN (binary)

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

Error rate and bounds

If $|\mathcal{T}| \to \infty \Rightarrow P(\omega_i|\mathbf{x}_N) \approx P(\omega_i|\mathbf{x})$

$$\begin{aligned} R_{\text{NN}}(\mathbf{x}) &= P(\omega_1|\mathbf{x}) \cdot P(\omega_2|\mathbf{x}_N) + P(\omega_2|\mathbf{x}) \cdot P(\omega_1|\mathbf{x}_N) \approx \\ &= P(\omega_1|\mathbf{x}) \cdot P(\omega_2|\mathbf{x}) + P(\omega_2|\mathbf{x}) \cdot P(\omega_1|\mathbf{x}) = \\ &= 2 \cdot P(\omega_1|\mathbf{x}) \cdot P(\omega_2|\mathbf{x}) = 2 \cdot P(\omega_1|\mathbf{x}) \left[1 - P(\omega_1|\mathbf{x})\right] = \\ &= 2 \cdot R^*(\mathbf{x}) \left[1 - R^*(\mathbf{x})\right] \end{aligned}$$

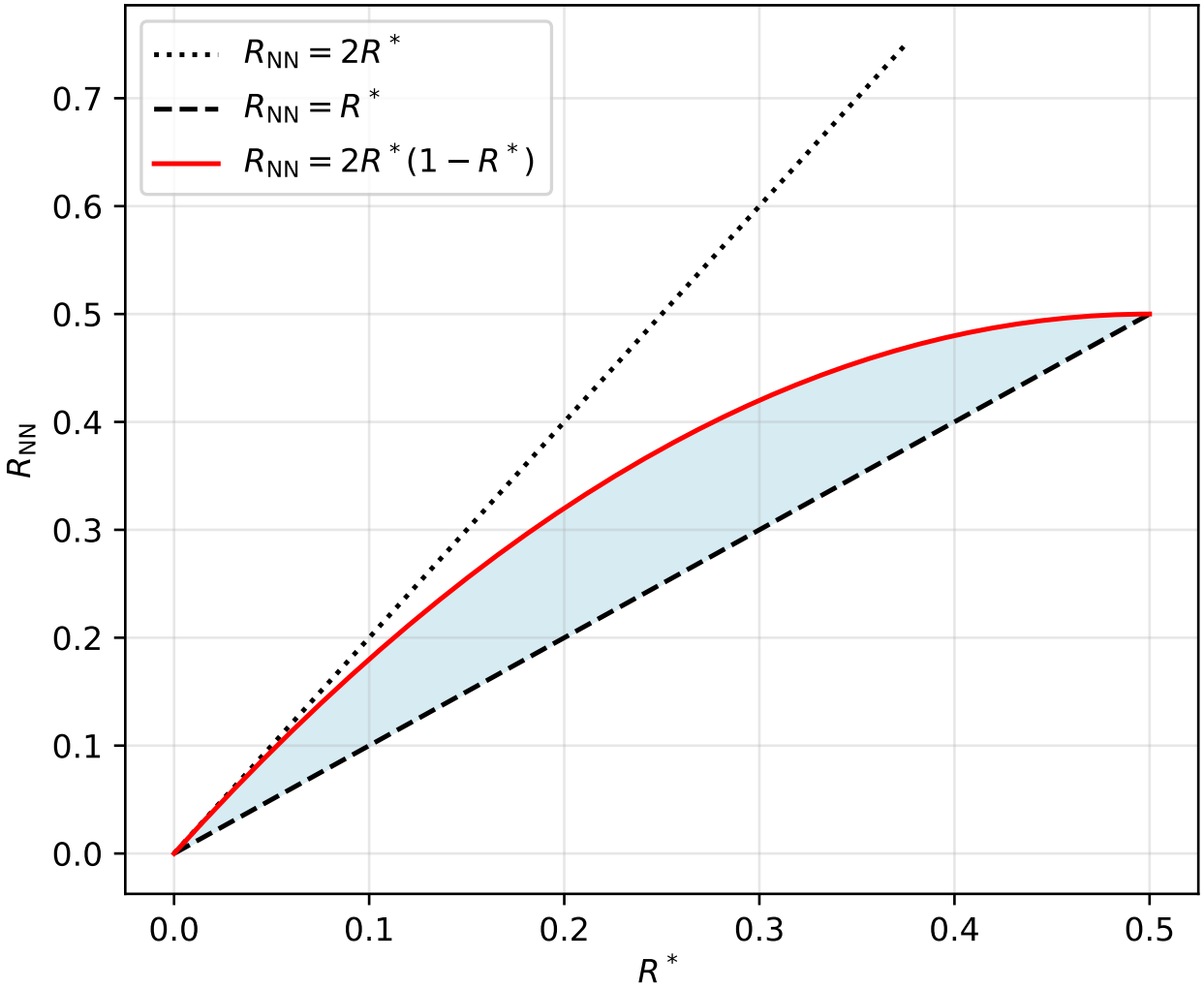
Error bound for NN (binary)

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

Error bound for NN (multiclass)

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

Error rate and bounds



Concept of metric

Concept of metric

- NN relies on a metric or distance function $D(\cdot, \cdot)$
  - $\Rightarrow$ Gives a generalized scalar distance between two arguments

Concept of metric

- NN relies on a metric or distance function $D(\cdot, \cdot)$
  - $\Rightarrow$ Gives a generalized scalar distance between two arguments

- A metric must follow four properties:
  1. Nonnegativity: $D(\mathbf{a}, \mathbf{b}) \geq 0$
  2. Reflexivity: $D(\mathbf{a}, \mathbf{b}) = 0$ iif $\mathbf{a} = \mathbf{b}$
  3. Symmetry: $D(\mathbf{a}, \mathbf{b}) = D(\mathbf{b}, \mathbf{a})$
  4. Triangle inequality: $D(\mathbf{a}, \mathbf{b}) + D(\mathbf{b}, \mathbf{c}) \geq D(\mathbf{a}, \mathbf{c})$

Minkowski distance

Generalized metric that unifies most well-known distance measures:

Minkowski distance

Generalized metric that unifies most well-known distance measures:

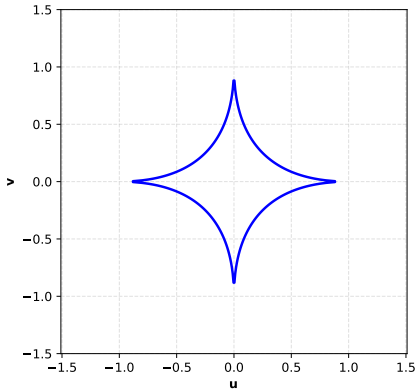
$$D(\mathbf{a}, \mathbf{b}) = \left(\sum_{i=1}^{|\mathbf{a}|} |a_i - b_i|^p \right)^{\frac{1}{p}} \quad \text{with } p \geq 1$$

Minkowski distance

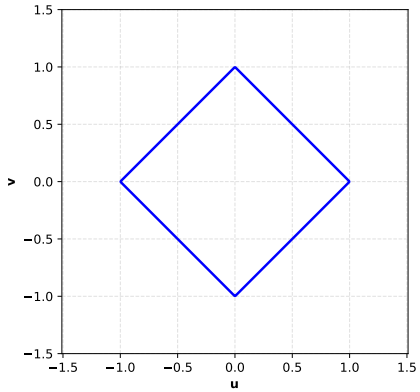
Generalized metric that unifies most well-known distance measures:

$$D(\mathbf{a}, \mathbf{b}) = \left(\sum_{i=1}^{|\mathbf{a}|} |a_i - b_i|^p \right)^{\frac{1}{p}} \quad \text{with } p \geq 1$$

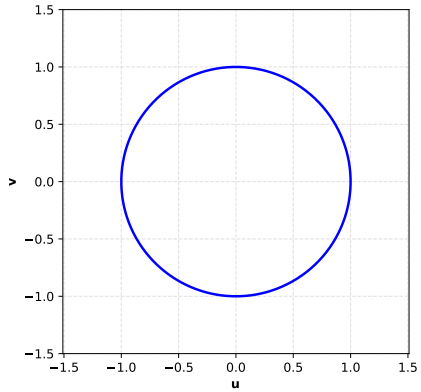
$p = 0.5$



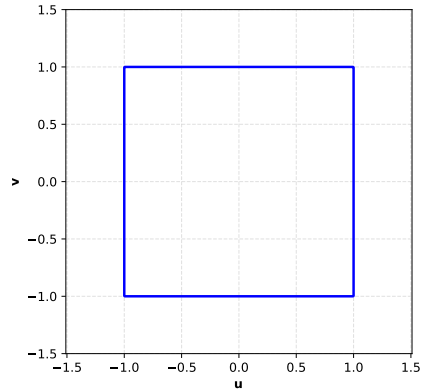
$p = 1$



$p = 2$



$p = \infty$



Formulation

Generalization of the NN rule $\Rightarrow$ considers the k nearest elements

Formulation

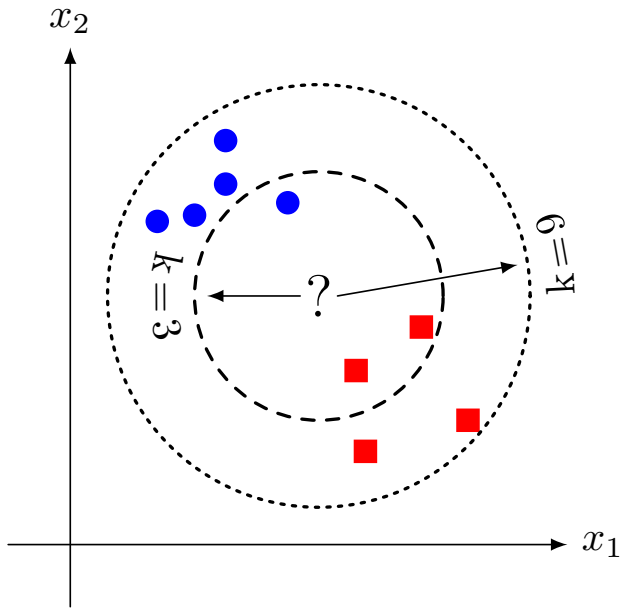
Generalization of the NN rule $\Rightarrow$ considers the k nearest elements

- Estimated class $\hat{\omega} \Rightarrow$ The most frequent label among the k neighbors

Formulation

Generalization of the NN rule $\Rightarrow$ considers the k nearest elements

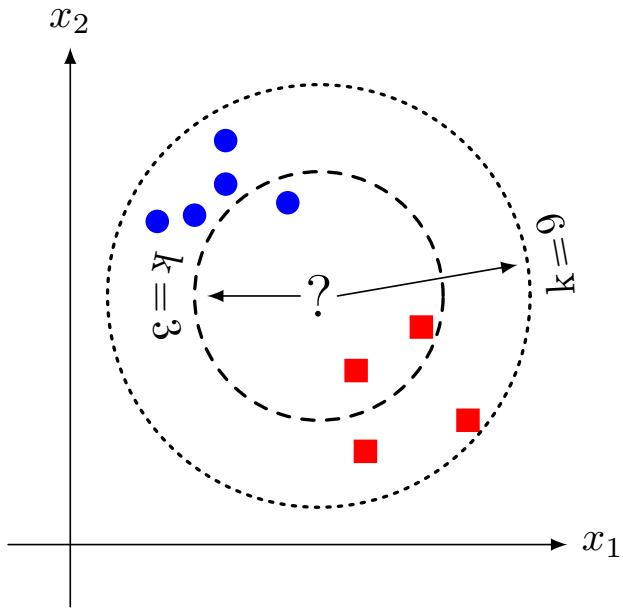
- Estimated class $\hat{\omega} \Rightarrow$ The most frequent label among the k neighbors



Formulation

Generalization of the NN rule $\Rightarrow$ considers the k nearest elements

- Estimated class $\hat{\omega} \Rightarrow$ The most frequent label among the k neighbors



Hyperparameter k:

- Low k values: Sensitive to noise (high variance)
- High k values: Too smooth/general (high bias)

Error rates

kNN with $k > 1$ trades bias for lower variance

Error rates

kNN with $k > 1$ trades bias for lower variance

$\rightarrow$ Estimated risk ($R_{k\text{NN}}$) tends to R^*

Error rates

kNN with $k > 1$ trades bias for lower variance

$\rightarrow$ Estimated risk ($R_{k\text{NN}}$) tends to R^*

Error bound for kNN (multiclass)

$$R^* \leq R_{k\text{NN}} \leq R_{\text{NN}} \leq R^* \left(2 - \frac{c}{c-1} R^*\right)$$

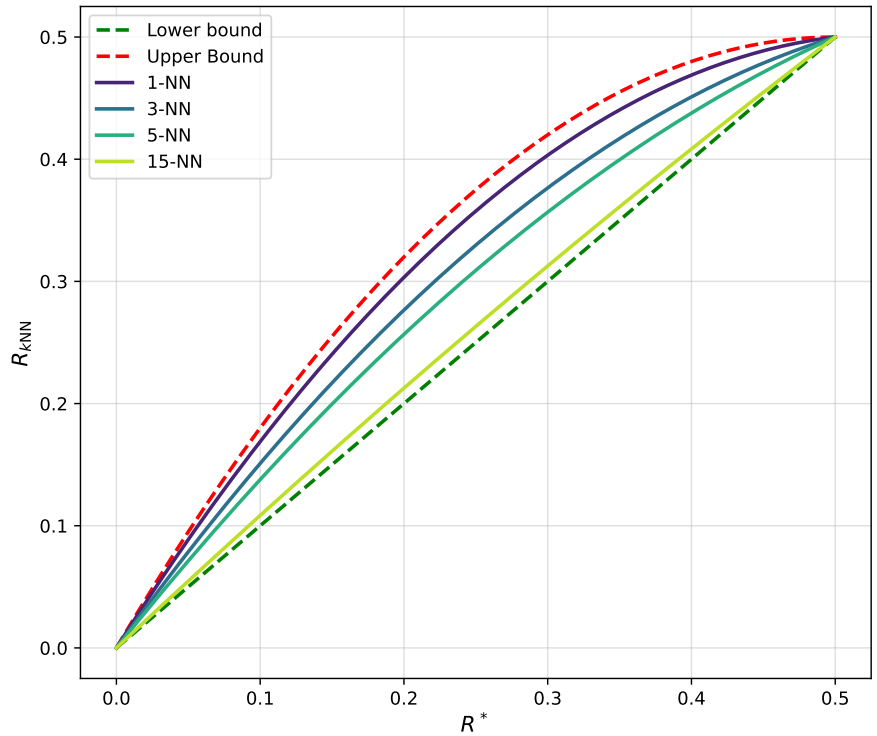
Error rates

kNN with $k > 1$ trades bias for lower variance

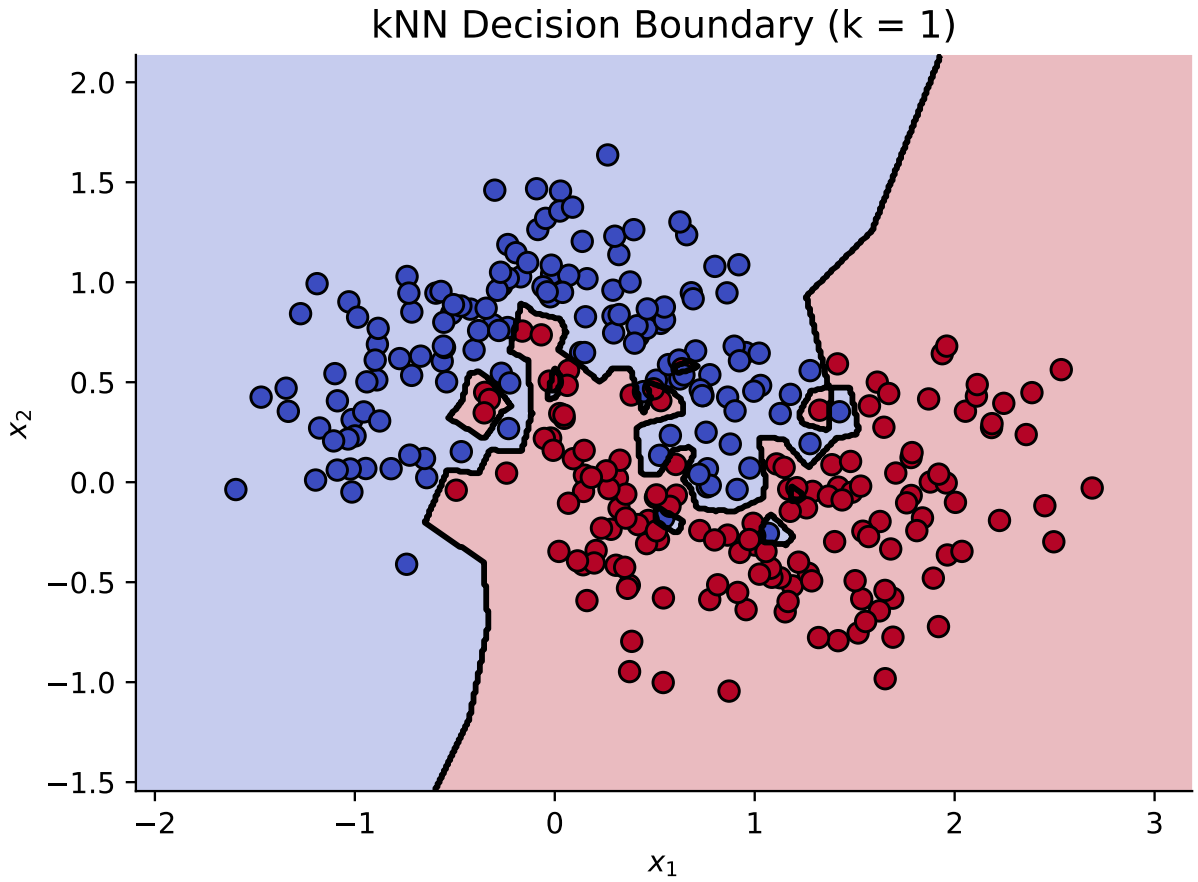
$\rightarrow$ Estimated risk ($R_{k\text{NN}}$) tends to R^*

Error bound for kNN (multiclass)

$$R^* \leq R_{k\text{NN}} \leq R_{\text{NN}} \leq R^* \left(2 - \frac{c}{c-1} R^*\right)$$



Decision boundaries



Practicalities

Practicalities

✓ Interactive learning / fast adaptation

- Model is updated by simply adding new elements

Practicalities

- ✓ Interactive learning / fast adaptation
  - Model is updated by simply adding new elements
- ✓ Good error properties
  - Bounded by twice the Bayes risk

Practicalities

- ✓ Interactive learning / fast adaptation
  - Model is updated by simply adding new elements
- ✓ Good error properties
  - Bounded by twice the Bayes risk
- ✓ Tasks beyond classification and regression
  - Search of elements in databases

Practicalities

- ✓ Interactive learning / fast adaptation
  - Model is updated by simply adding new elements
- ✓ Good error properties
  - Bounded by twice the Bayes risk
- ✓ Tasks beyond classification and regression
  - Search of elements in databases
- ✗ Features may depict different ranges
  - Dominant feature that eclipses the others $\Rightarrow$ Normalization

Practicalities

- ✓ Interactive learning / fast adaptation
  - Model is updated by simply adding new elements
- ✓ Good error properties
  - Bounded by twice the Bayes risk
- ✓ Tasks beyond classification and regression
  - Search of elements in databases
- ✗ Features may depict different ranges
  - Dominant feature that eclipses the others $\Rightarrow$ Normalization
- ✗ Inefficient classifier $\Rightarrow \mathcal{O}(n^2)$
  1. Fast similarity search
  2. Approximate search
  3. Data reduction

Outline

Contextualization

- k-**Nearest Neighbor**: relevant example of nonparametric learning
  $\Rightarrow$ But there exist other types of models

Contextualization

- k-**Nearest Neighbor**: relevant example of nonparametric learning
  $\Rightarrow$ But there exist other types of models

- Each model has its own idiosyncratic way of dividing the label space:
  $\Rightarrow$ Different types of decision boundaries

Contextualization

- k-**Nearest Neighbor**: relevant example of nonparametric learning
  $\Rightarrow$ But there exist other types of models

- Each model has its own idiosyncratic way of dividing the label space:
  $\Rightarrow$ Different types of decision boundaries

- Two examples of models:
  1. Decision tree
  2. Support Vector Machine

Decision tree

Decision tree

- Derives a tree structure:
  - → Nodes evaluate the value of the feature
  - → Leaves denote the labels

Decision tree

- Derives a tree structure:
  - → Nodes evaluate the value of the feature
  - → Leaves denote the labels

- Typically constructed following information gain principles

Decision tree

- Derives a tree structure:
  - → Nodes evaluate the value of the feature
  - → Leaves denote the labels

- Typically constructed following information gain principles

- Tend to overfit:
  - ✗ Pruning strategies to improve generalization
  - ✓ Useful for outlier detection

Decision tree

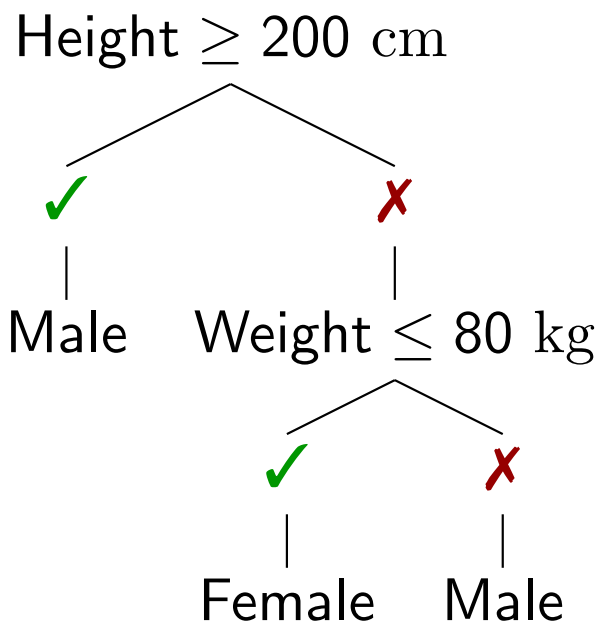
Example:

- Binary task: $\mathcal{W} = \{\text{male}, \text{female}\}$
- Feature space $\mathbb{R}^2$: weight (kg), height (cm)

Decision tree

Example:

- Binary task: $\mathcal{W} = \{\text{male}, \text{female}\}$
- Feature space $\mathbb{R}^2$: weight (kg), height (cm)



Support Vector Machine

Support Vector Machine

- Linear binary classifier that uses a hyperplane to separate the labels
  → Support vectors: elements that define the hyperplane

Support Vector Machine

- Linear binary classifier that uses a hyperplane to separate the labels
  → Support vectors: elements that define the hyperplane

- Kernel: maps the data into a higher dimensionality space

Support Vector Machine

- Linear binary classifier that uses a hyperplane to separate the labels
  → Support vectors: elements that define the hyperplane

- Kernel: maps the data into a higher dimensionality space

- Distinct classes must be in different *sides* of the hyperplane
  → Premise is relaxed via *penalties*

Support Vector Machine

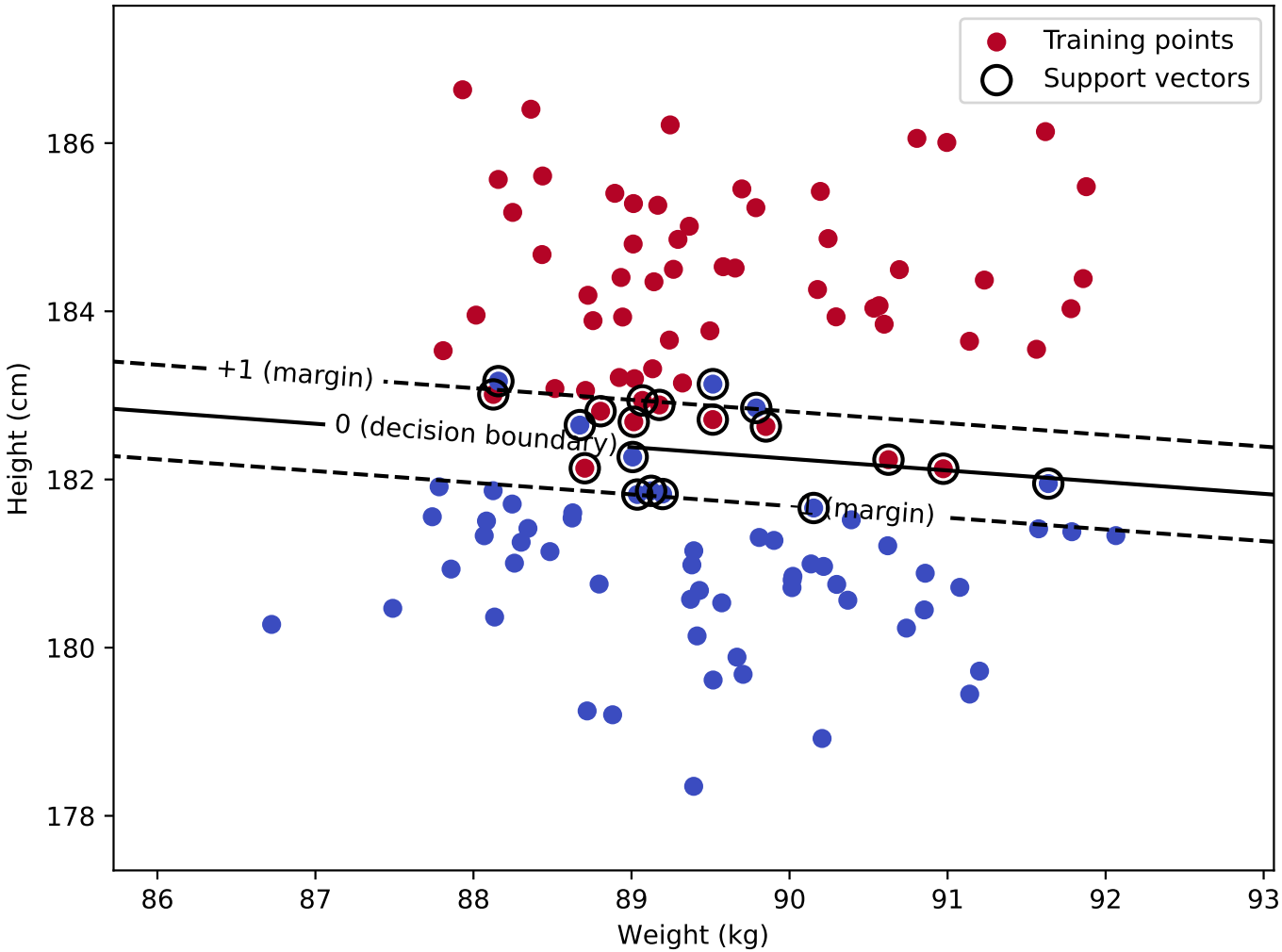
- Linear binary classifier that uses a hyperplane to separate the labels
  → Support vectors: elements that define the hyperplane

- Kernel: maps the data into a higher dimensionality space

- Distinct classes must be in different *sides* of the hyperplane
  → Premise is relaxed via *penalties*

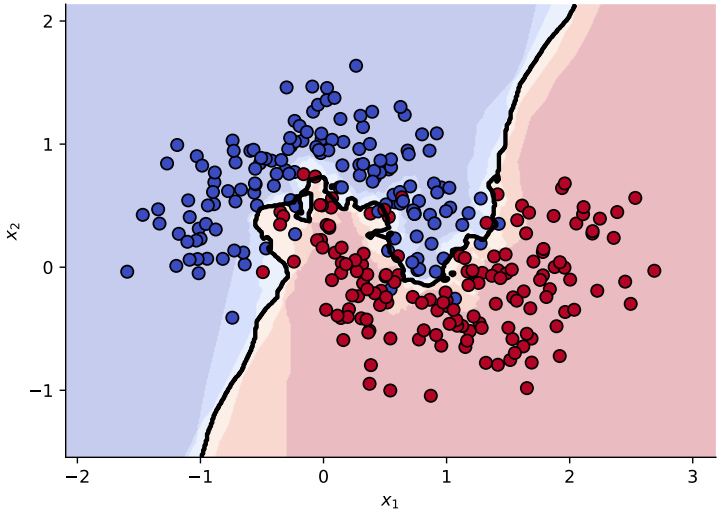
- Adaptations for multiclass scenarios
  → One-vs-all configuration
  → One-vs-one configuration

Support Vector Machine

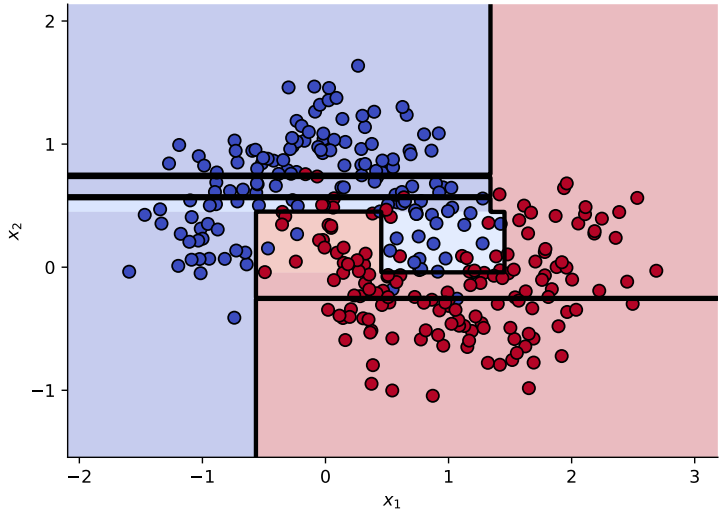


Decision boundaries (comparison)

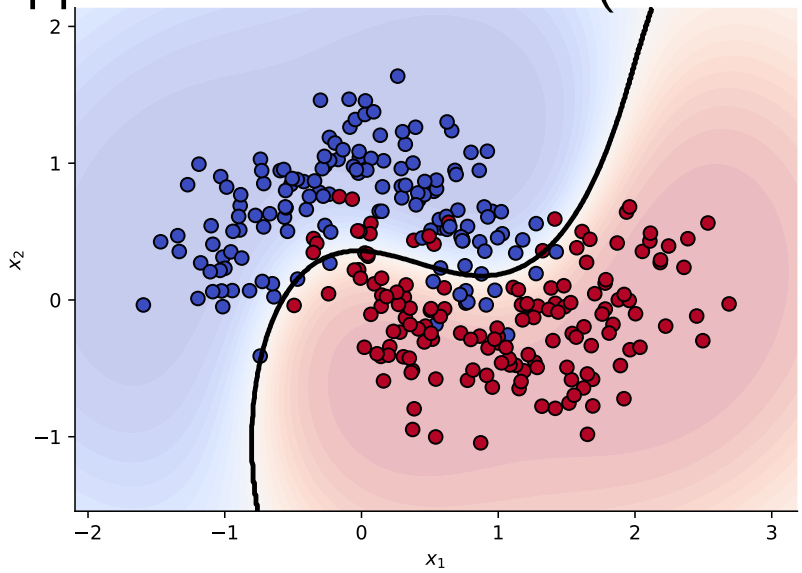
kNN with $k = 5$



Decision Tree with 5 levels



Support Vector Machine (RBF kernel)



T4: Nonparametric and distance-based learning

Fundamentos del Aprendizaje Automático

Curso 2025/2026