Machine Learning I
80-629A

Apprentissage Automatique I
80-629

Supervised Learning
— Week #3

Today: Models for supervised learning

- (Mostly) linear models
- Focus on classification

1. Non-Probabilistic Models
    - Nearest Neighbor, Support Vector Machines (SVMs)
2. Probabilistic Models
    - Naive Bayes

Supervised learning

Train Data

$$\underbrace{\begin{array}{c|ccccc|}
 & \text{Nb.bed.} & \text{Area} & \text{Neigh.} & \cdot & \cdot \\
x_0 & 1 & 0 & 0 & 0 & 0 \\
x_1 & 1 & 100 & 1 & .2 & .5 \\
x_2 & 3 & 200 & 0 & .1 & .2 \\
x_3 & 1 & 150 & 1 & .4 & .1 \\
x_4 & 2 & 210 & 2 & .5 & 1.1
\end{array}}_{X}
\quad
\underbrace{\begin{array}{c|c|}
 & \text{Sell-ability} \\
y_0 & 1 \\
y_1 & 2 \\
y_2 & 0 \\
y_3 & 2 \\
y_4 & 1
\end{array}}_{Y}$$

Task

$$f : \mathbb{R}^n \rightarrow \{0, 1, 2\}$$

Test Data

$$\underbrace{\begin{array}{c|ccccc|}
 & \text{Nb.bed.} & \text{Area} & \text{Neigh.} & \cdot & \cdot \\
x_0 & 1 & 0 & 0 & 0 & 0 \\
x_1 & 2 & 50 & 1 & .3 & .8 \\
x_2 & 1 & 100 & 1 & .5 & 1.4 \\
x_3 & 4 & 170 & 0 & .7 & .4 \\
x_4 & 1 & 120 & 3 & .9 & .5
\end{array}}_{X^{new}}
\quad
\underbrace{\begin{array}{c|c|}
 & \text{Sell-ability} \\
y_0 & ? \\
y_1 & ? \\
y_2 & ? \\
y_3 & ? \\
y_4 & ?
\end{array}}_{Y^{new}}$$

Supervised learning

Train Data

$$\underbrace{\begin{array}{c} \\ x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{array}\begin{array}{c} \begin{array}{ccccc} \text{Nb.bed.} & \text{Area} & \text{Neigh.} & \cdot & \cdot \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 100 & 1 & .2 & .5 \\ 3 & 200 & 0 & .1 & .2 \\ 1 & 150 & 1 & .4 & .1 \\ 2 & 210 & 2 & .5 & 1.1 \end{bmatrix} \end{array}}_{X} \quad \underbrace{\begin{array}{c} \\ y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \end{array}\begin{array}{c} \text{Sell-ability} \\ \begin{bmatrix} 1 \\ 2 \\ 0 \\ 2 \\ 1 \end{bmatrix} \end{array}}_{Y}$$

Task

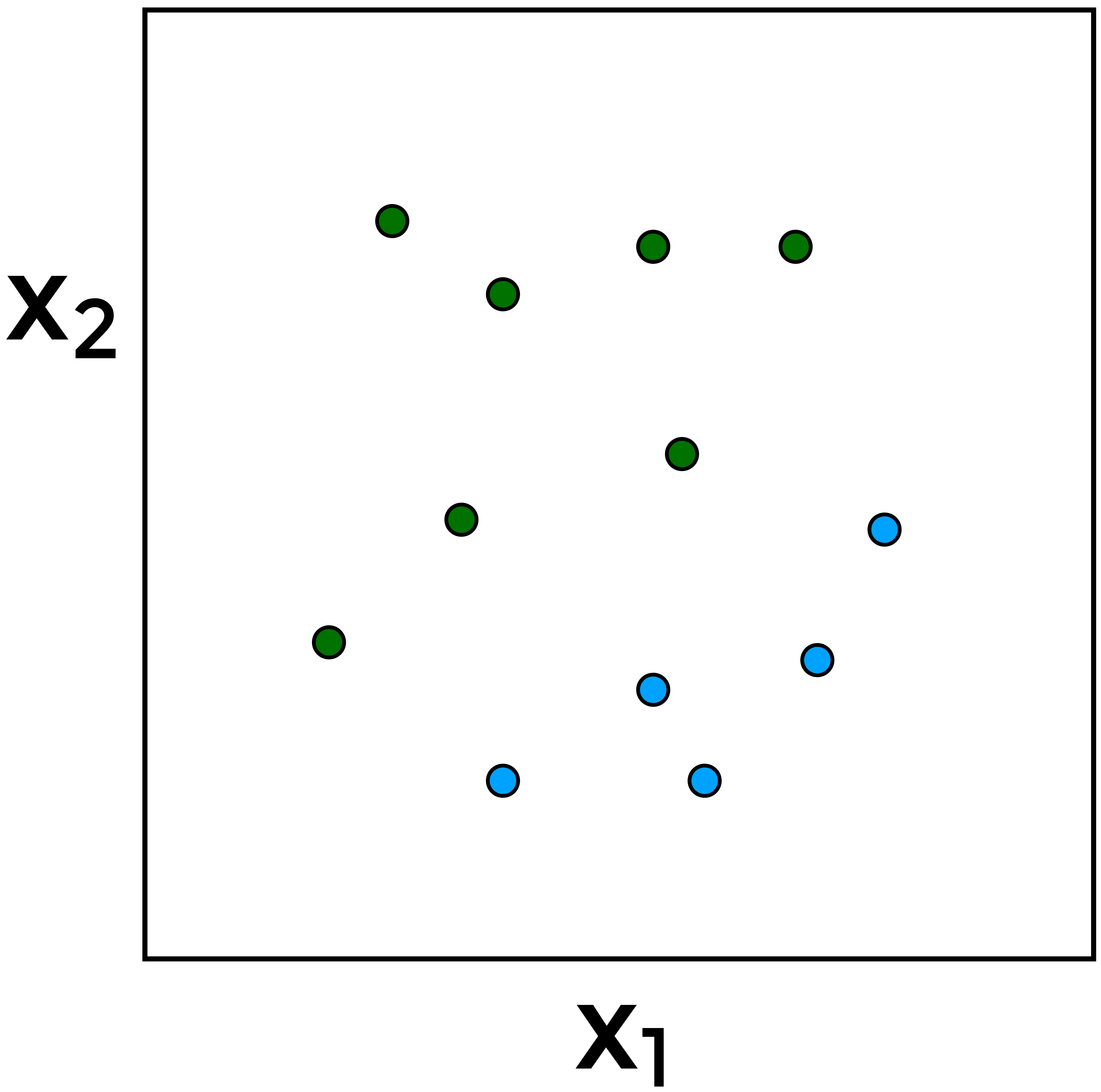
Models $\boxed{f}\ \mathbb{R}^n \to \{0, 1, 2\}$

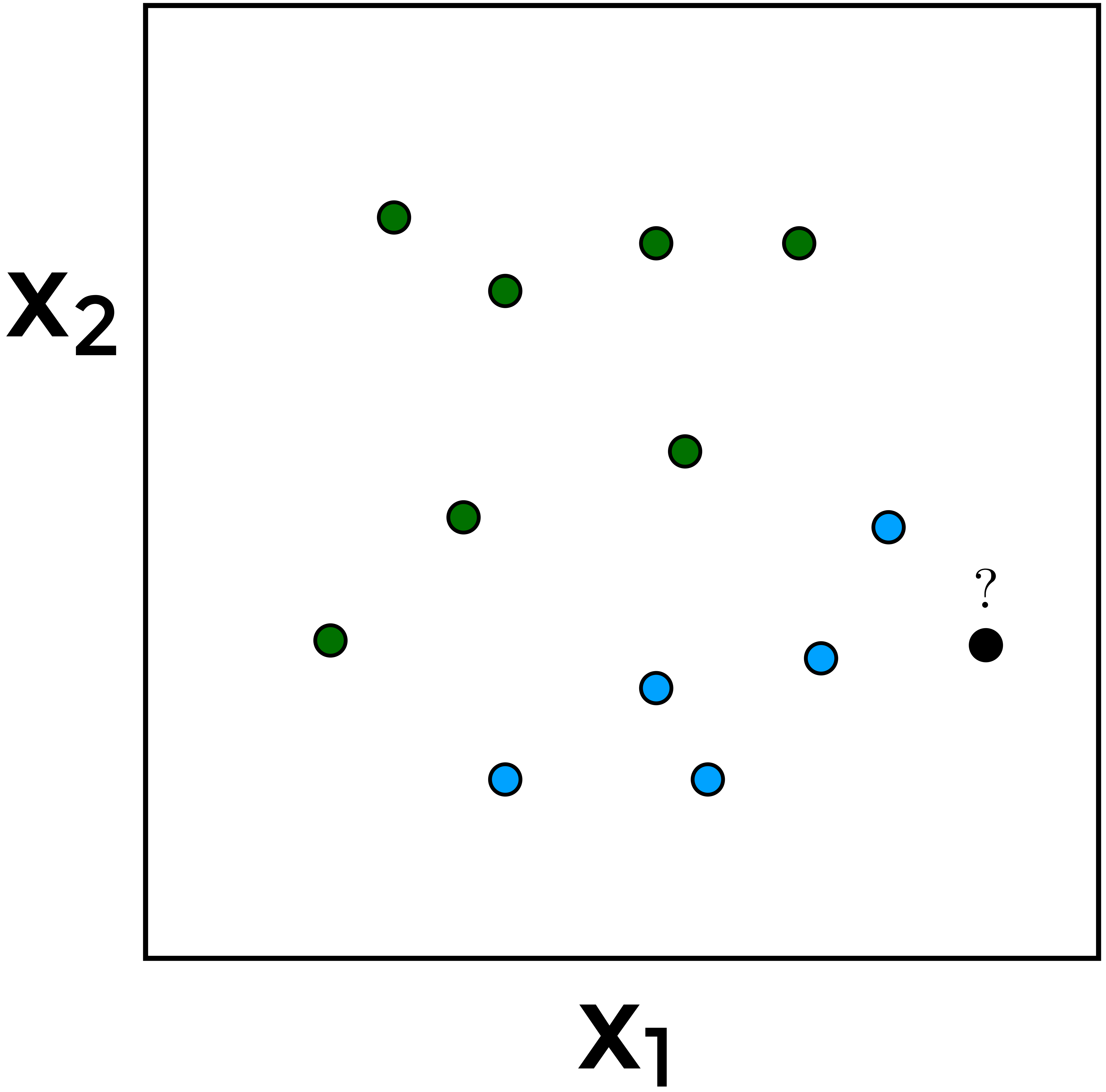
Test Data

$$\underbrace{\begin{array}{c} \\ x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{array}\begin{array}{c} \begin{array}{ccccc} \text{Nb.bed.} & \text{Area} & \text{Neigh.} & \cdot & \cdot \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 2 & 50 & 1 & .3 & .8 \\ 1 & 100 & 1 & .5 & 1.4 \\ 4 & 170 & 0 & .7 & .4 \\ 1 & 120 & 3 & .9 & .5 \end{bmatrix} \end{array}}_{X^{new}} \quad \underbrace{\begin{array}{c} \\ y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \end{array}\begin{array}{c} \text{Sell-ability} \\ \begin{bmatrix} ? \\ ? \\ ? \\ ? \\ ? \end{bmatrix} \end{array}}_{Y^{new}}$$

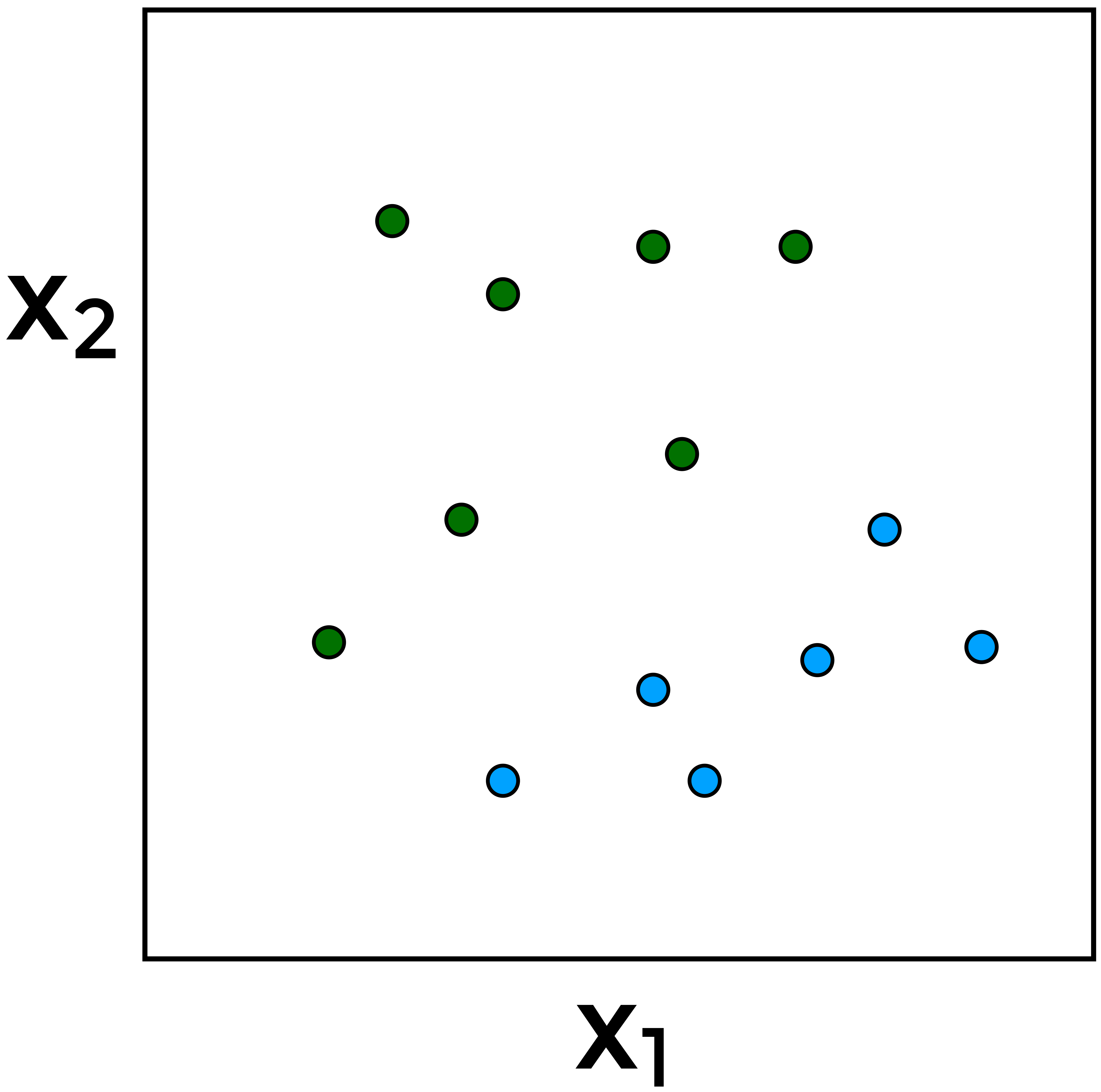
Nearest Neighbor (NN)

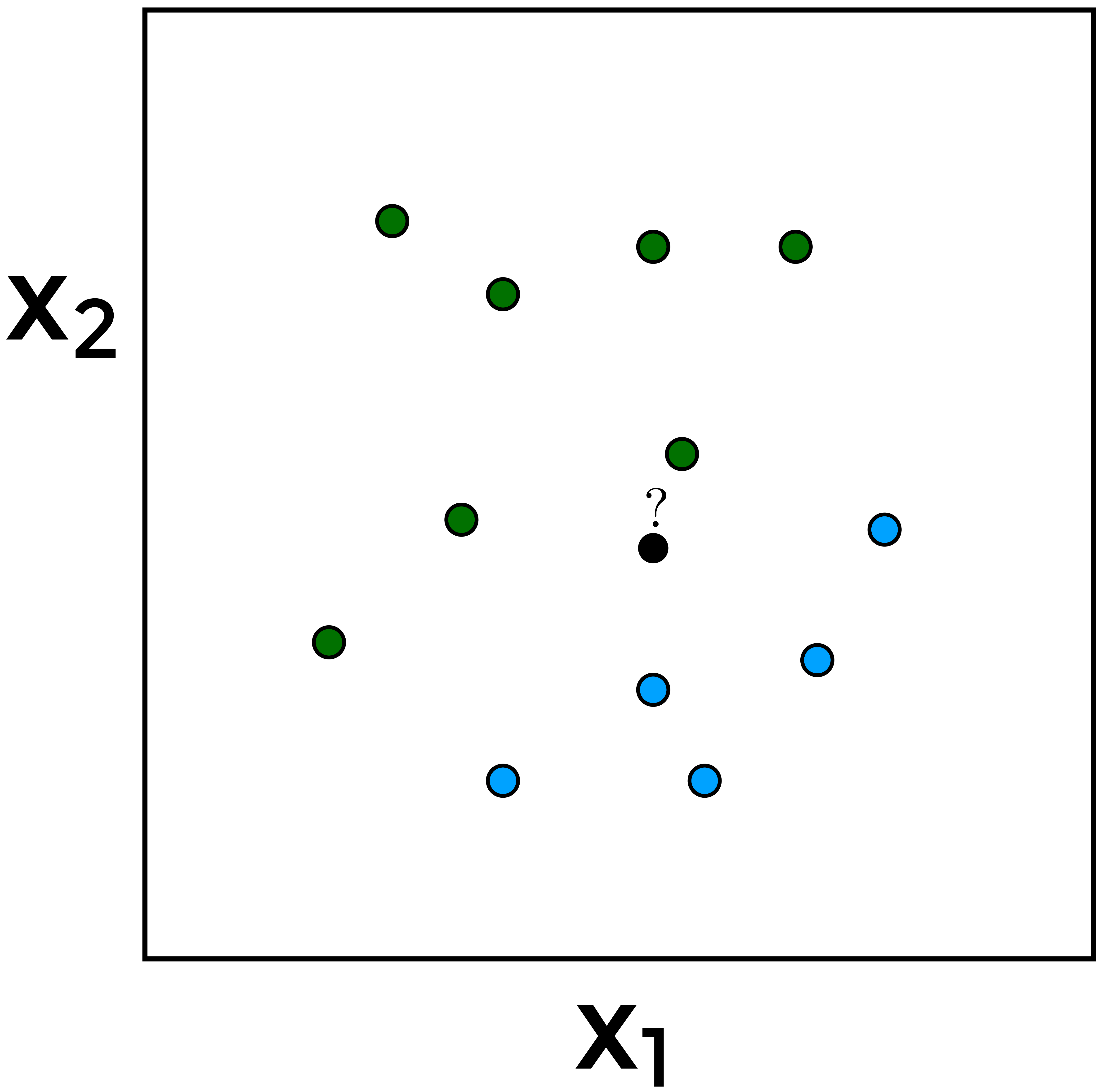
- Conceptually simple yet very powerful model
- Non-parametric model
  - No training phase
  - Test: Predict according to the neighbors of the instance

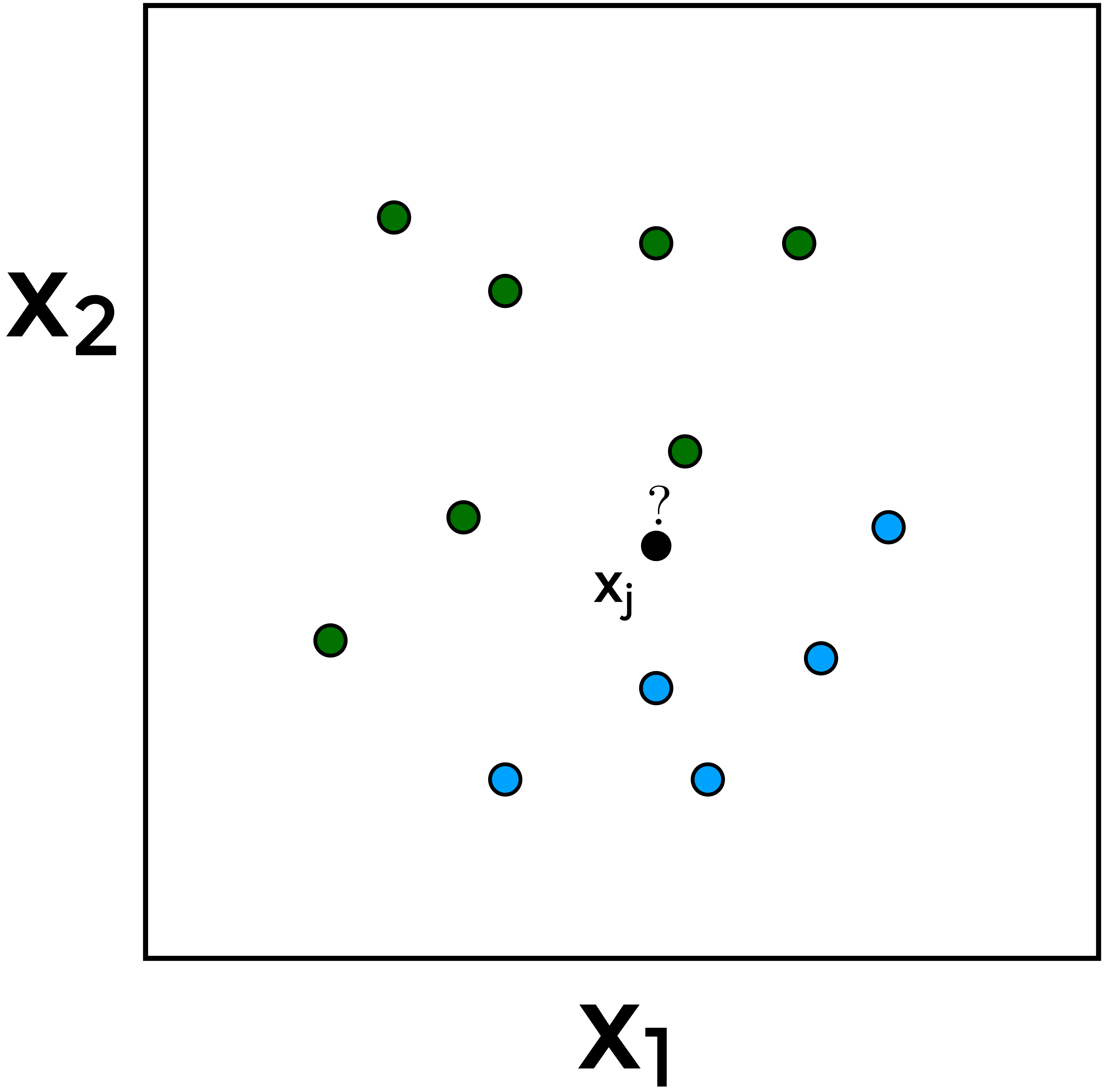

x_2
x_1



x_2
?
x_1


x_2
x_1

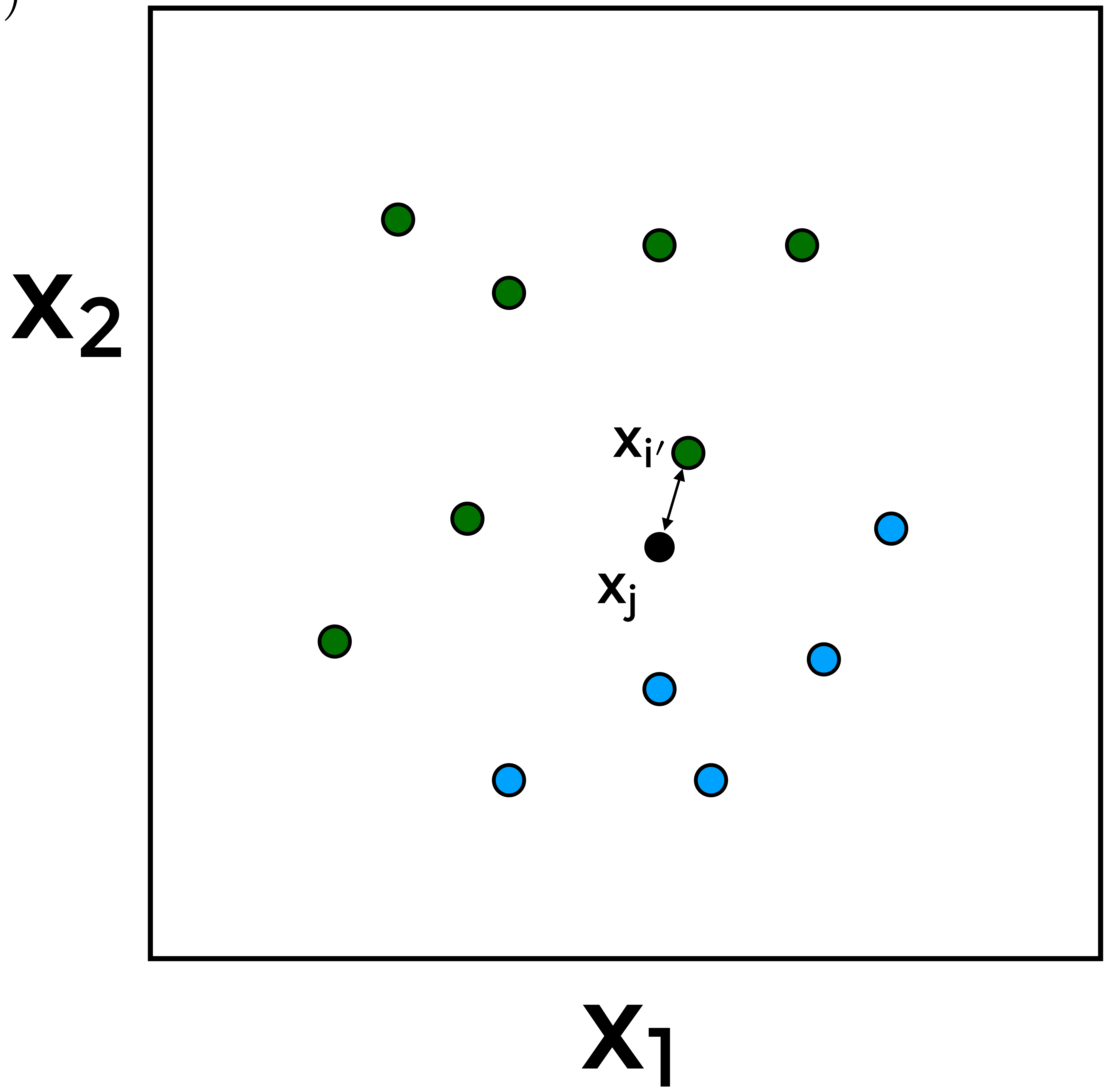

x2
?
x1



?
x_j
x_2
x_1

$$\mathsf{i}' = \arg\min_{\mathsf{i}} \mathsf{dist}(\mathsf{x_i}, \mathsf{x_j})$$

$$\mathsf{y_j} = \mathsf{y_{i'}}$$



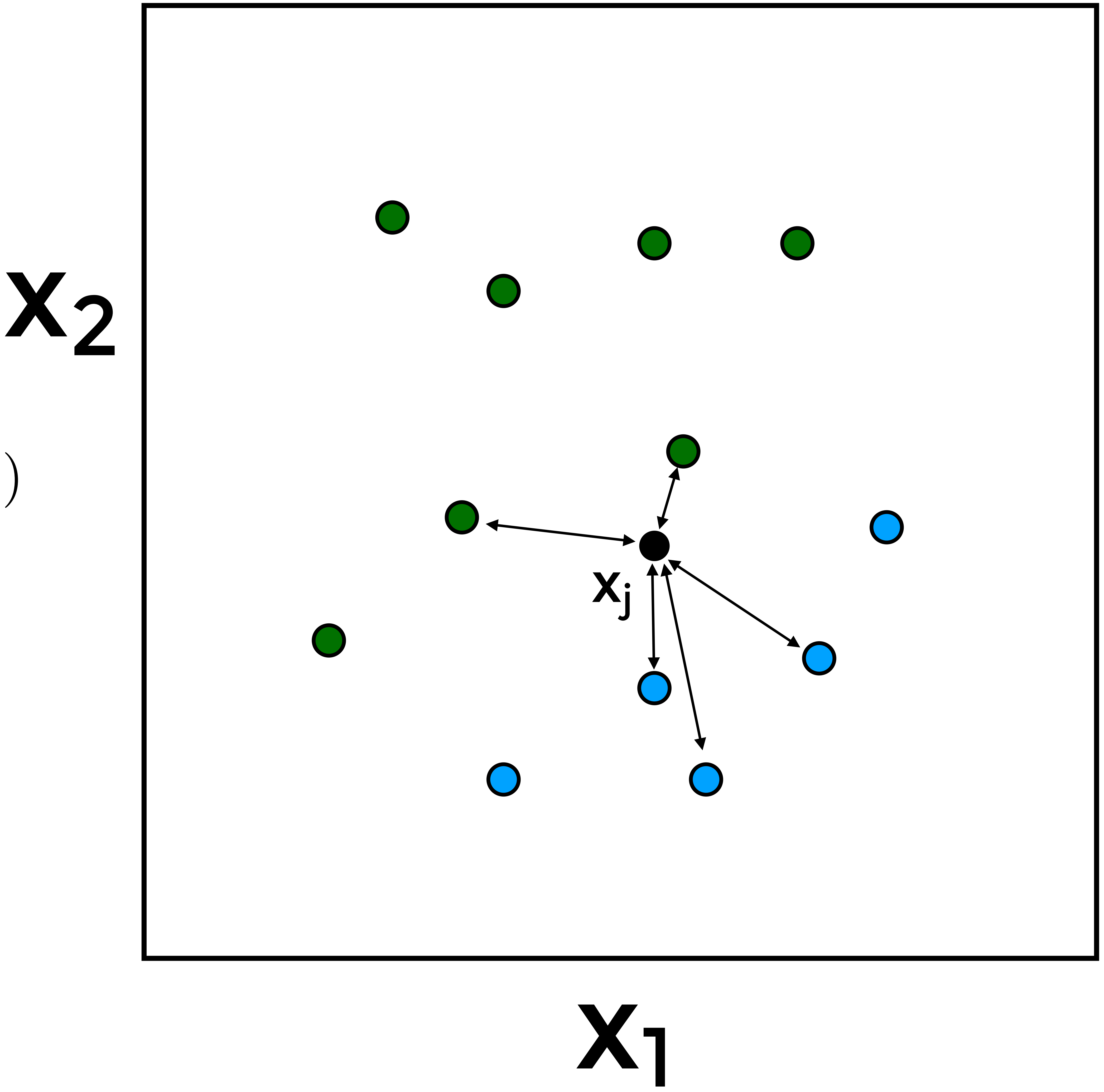
- **1-NN**

Instance classified according to its nearest neighbor

$\mathsf{k} = \mathsf{5}$ (assumption)

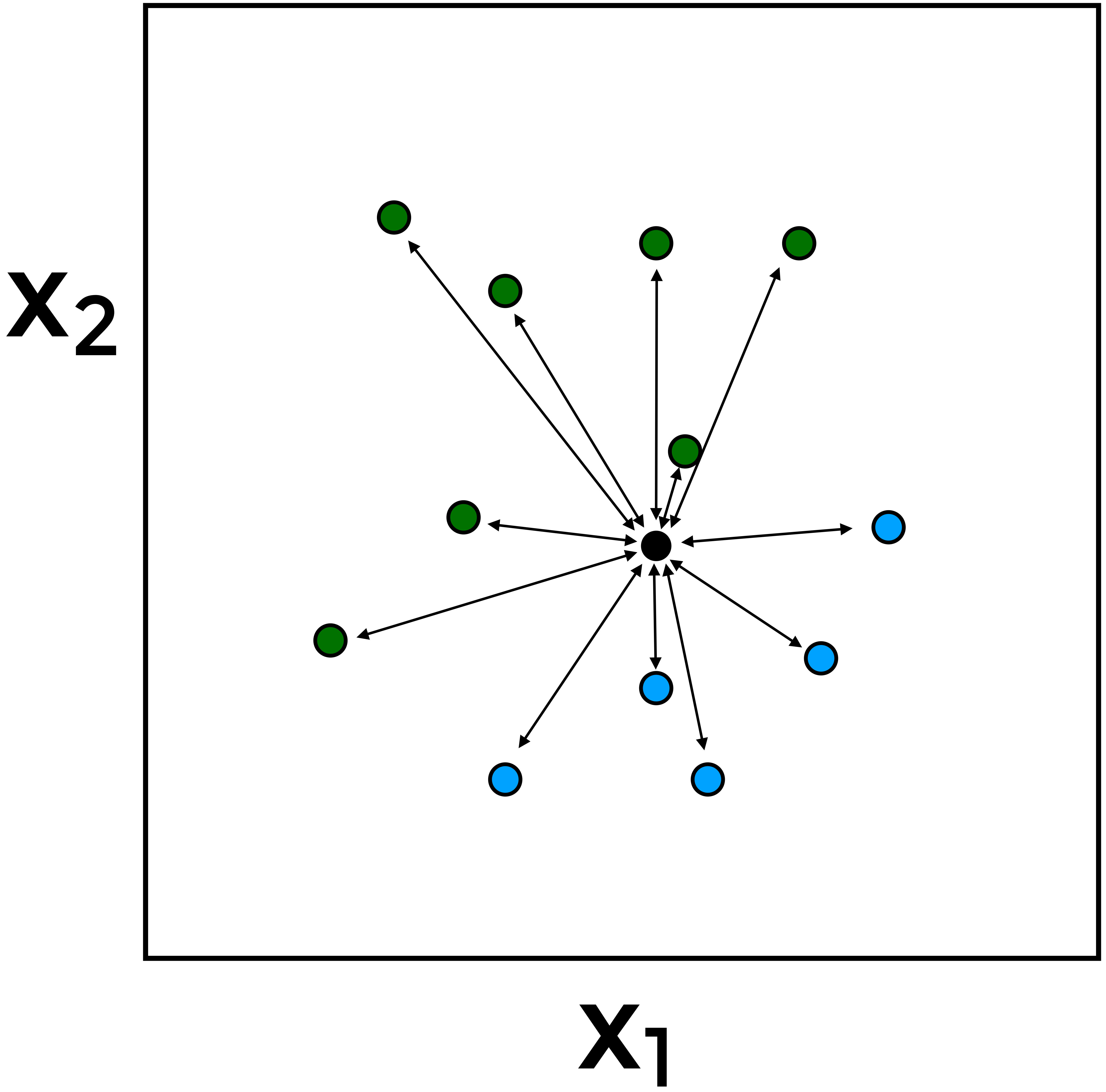
$\mathsf{i} = \arg\mathrm{sort}_{\mathsf{i}}\, \mathsf{dist}(\mathsf{x_i}, \mathsf{x_j})$

$\mathsf{y_j} = \mathsf{majority}(\mathsf{i}_{:5})$



- **K-NN**

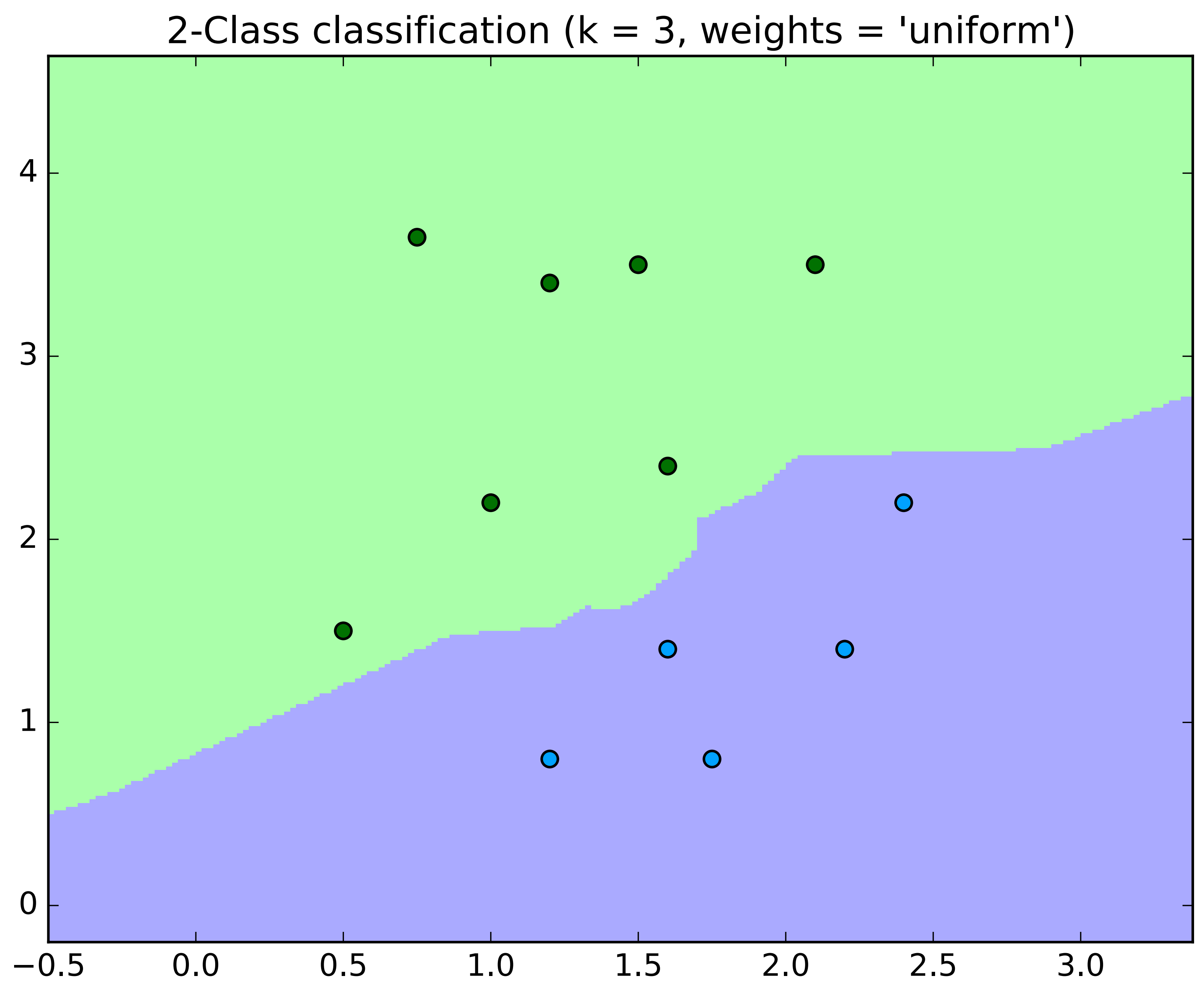
Instance classified according to the majority of its K nearest neighbors

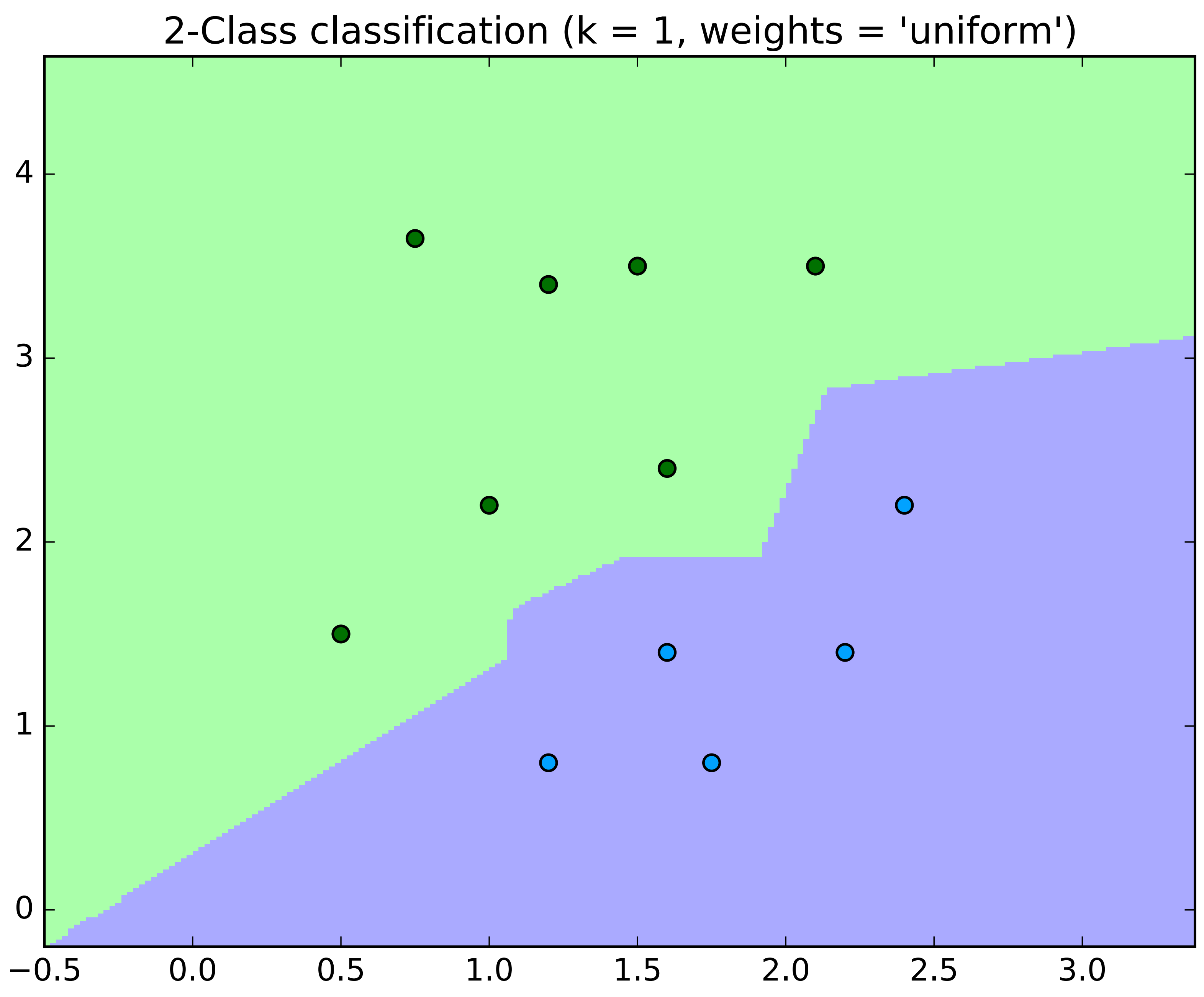


- **weighted-NN**

Instance classified according to all neighbors. The contribution of each neighbor is weighted by its distance.

Decision boundary

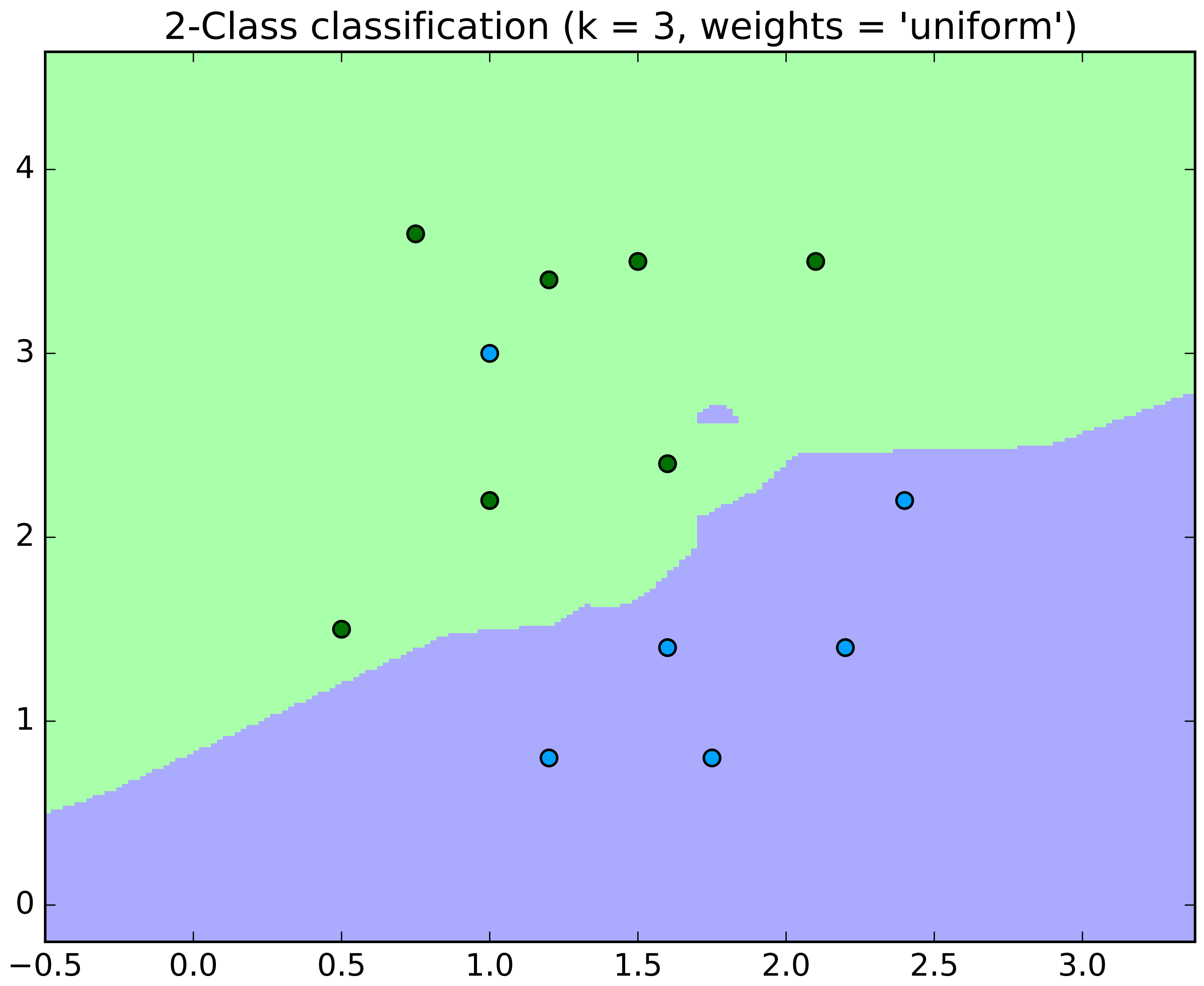


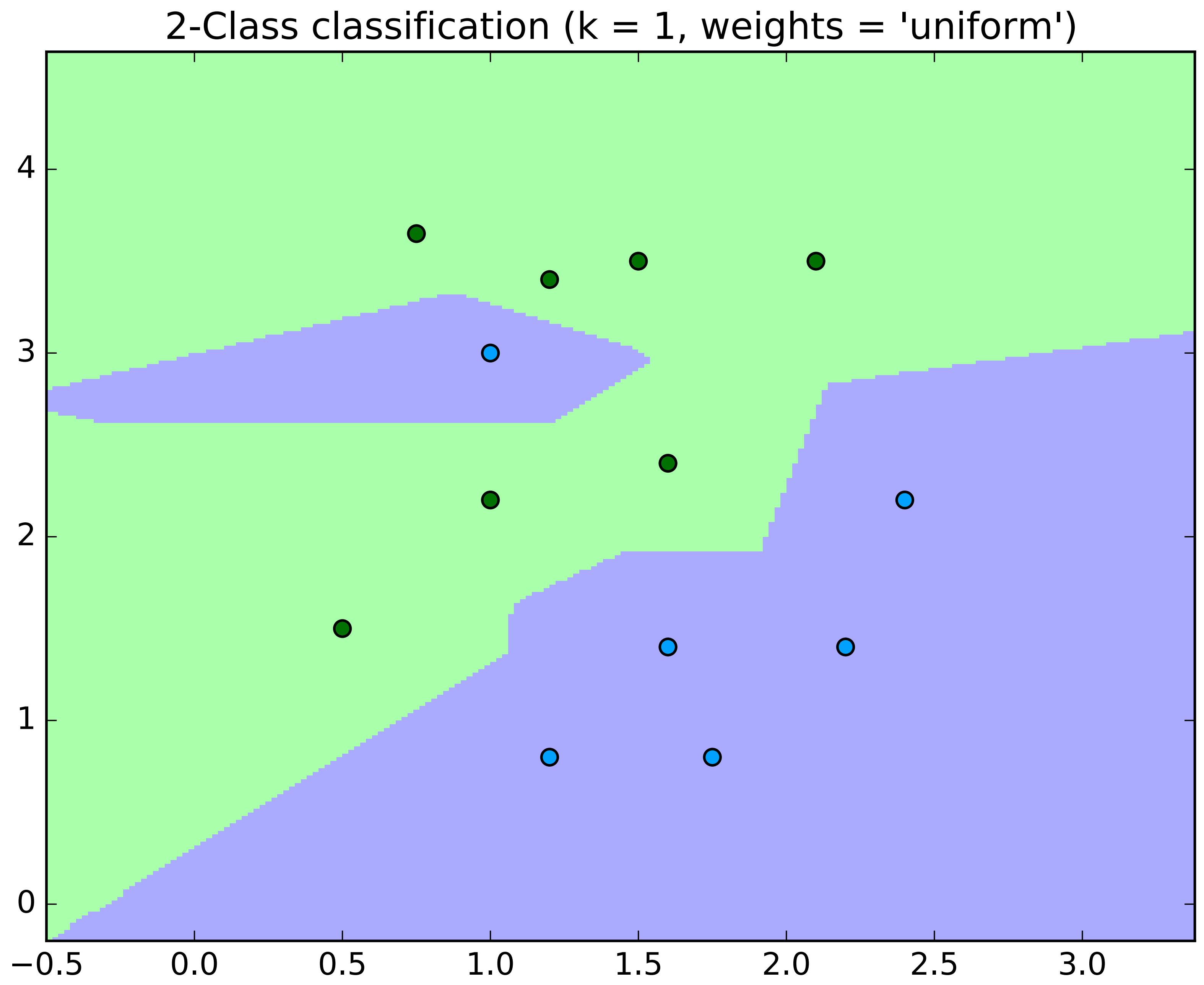



```
    # Scikit-learn library
    clf = neighbors.KNeighborsClassifier(n_neighbors, weights='uniform')
    clf.fit(X, y)
```

[Using: http://scikit-learn.org/stable/auto_examples/neighbors/plot_classification.html]

Adding a “noisy” instance





[Using: http://scikit-learn.org/stable/auto_examples/neighbors/plot_classification.html]

NN properties (I)

- Requires storing the whole dataset

  - Searching for nearest neighbour of a new datum can be expensive

    - scikit-learn has options for faster (approximate) searches

      algorithm : {'auto', 'ball_tree', 'kd_tree', 'brute'}, optional
      Algorithm used to compute the nearest neighbors:
      - 'ball_tree' will use `BallTree`
      - 'kd_tree' will use `KDTree`
      - 'brute' will use a brute-force search.
      - 'auto' will attempt to decide the most appropriate algorithm based on the values passed to `fit` method.

  - Can also work with non-continuous data

    p : integer, optional (default = 2)
    Power parameter for the Minkowski metric. When p = 1, this is equivalent to using manhattan_distance (l1), and euclidean_distance (l2) for p = 2.
    For arbitrary p, minkowski_distance (l_p) is used.

[http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html]

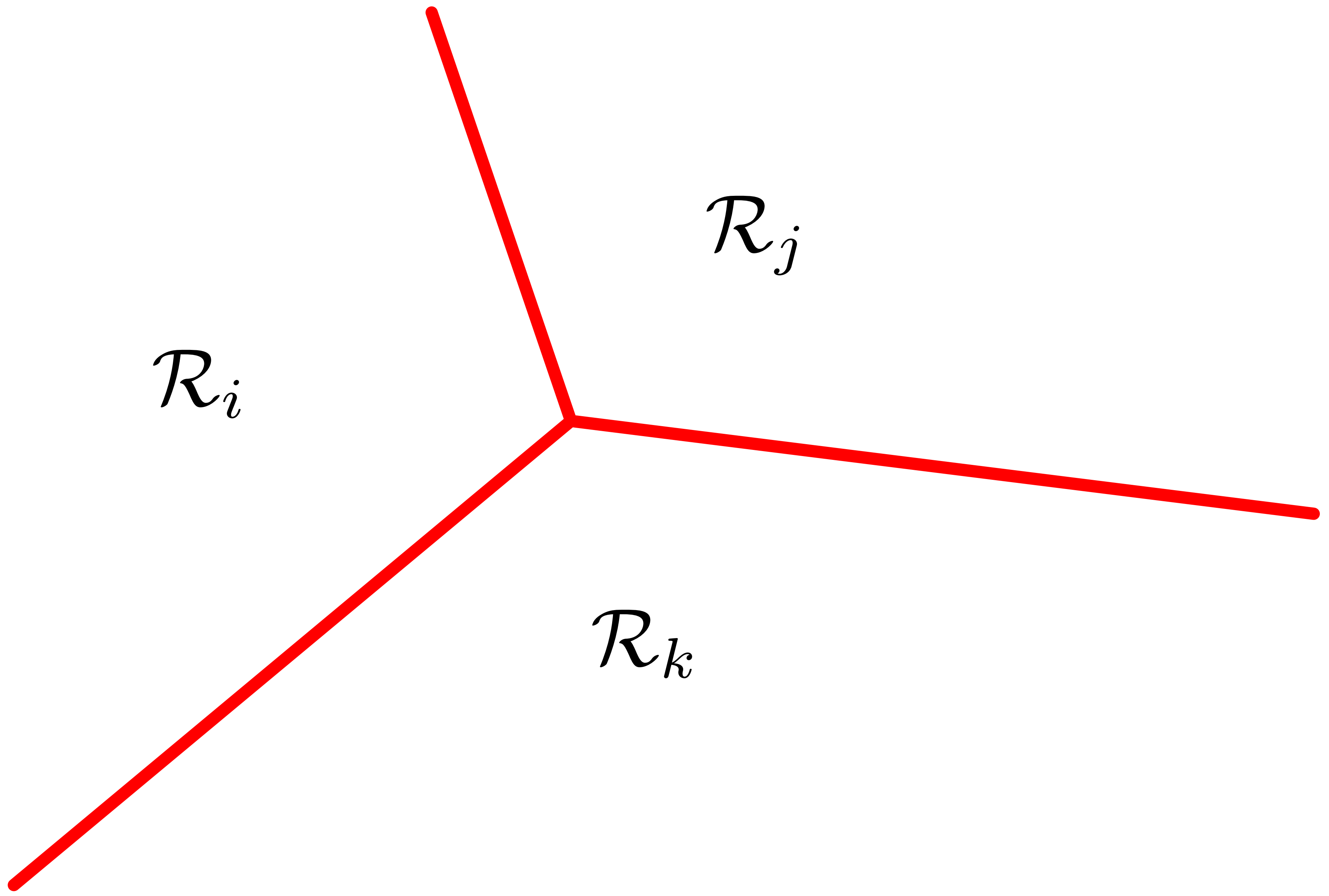
NN properties (II)

- In the limit of $N \to \infty$ the error rate is bounded by twice the optimal error (for K=1)
- May not perform well with high-dimensional inputs due to the **curse of dimensionality** (i.e., may use very far neighbours)

NN summary

- Non-parametric approach
  - Does not require fitting parameters
  - Hyper-parameter is the number of neighbors
- Also good for regression and density estimation

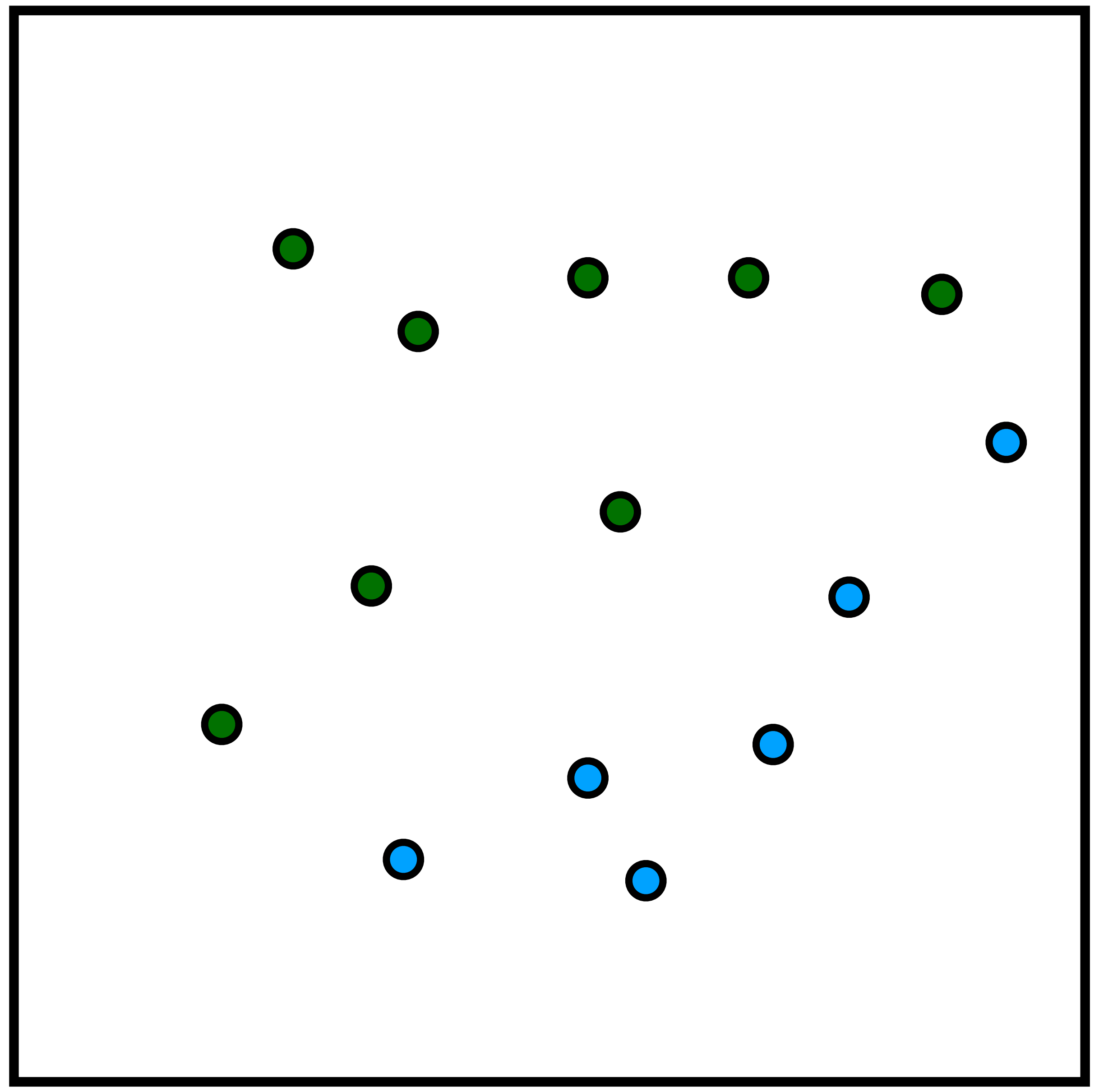
Linear Classification



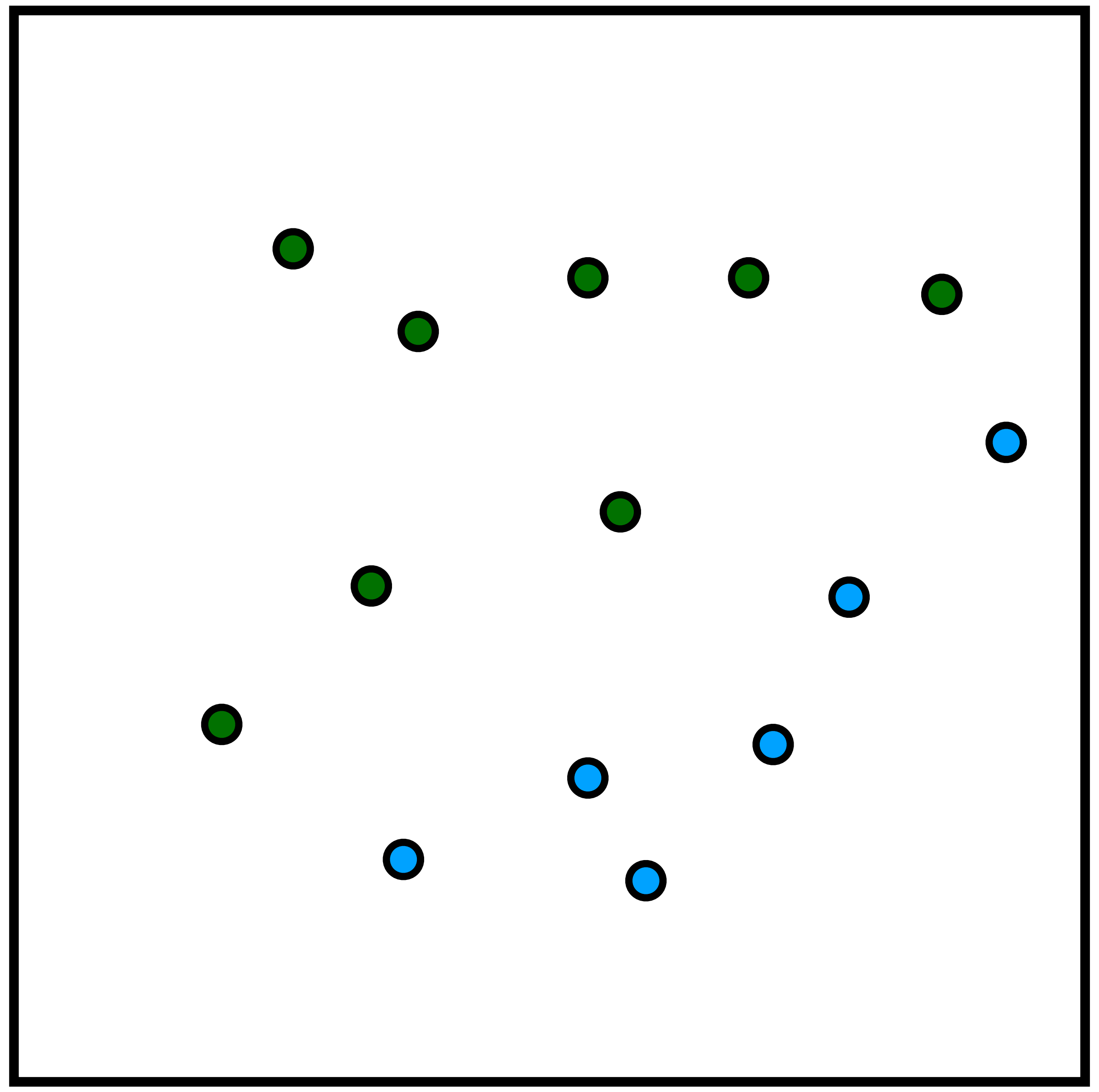
[Figure 4.3, Pattern Recognition & Machine Learning
C. Bishop]

- Divide the space into regions
- Different regions correspond to different predictions
- Frontiers between regions are called **decision boundaries**

Linear Classification

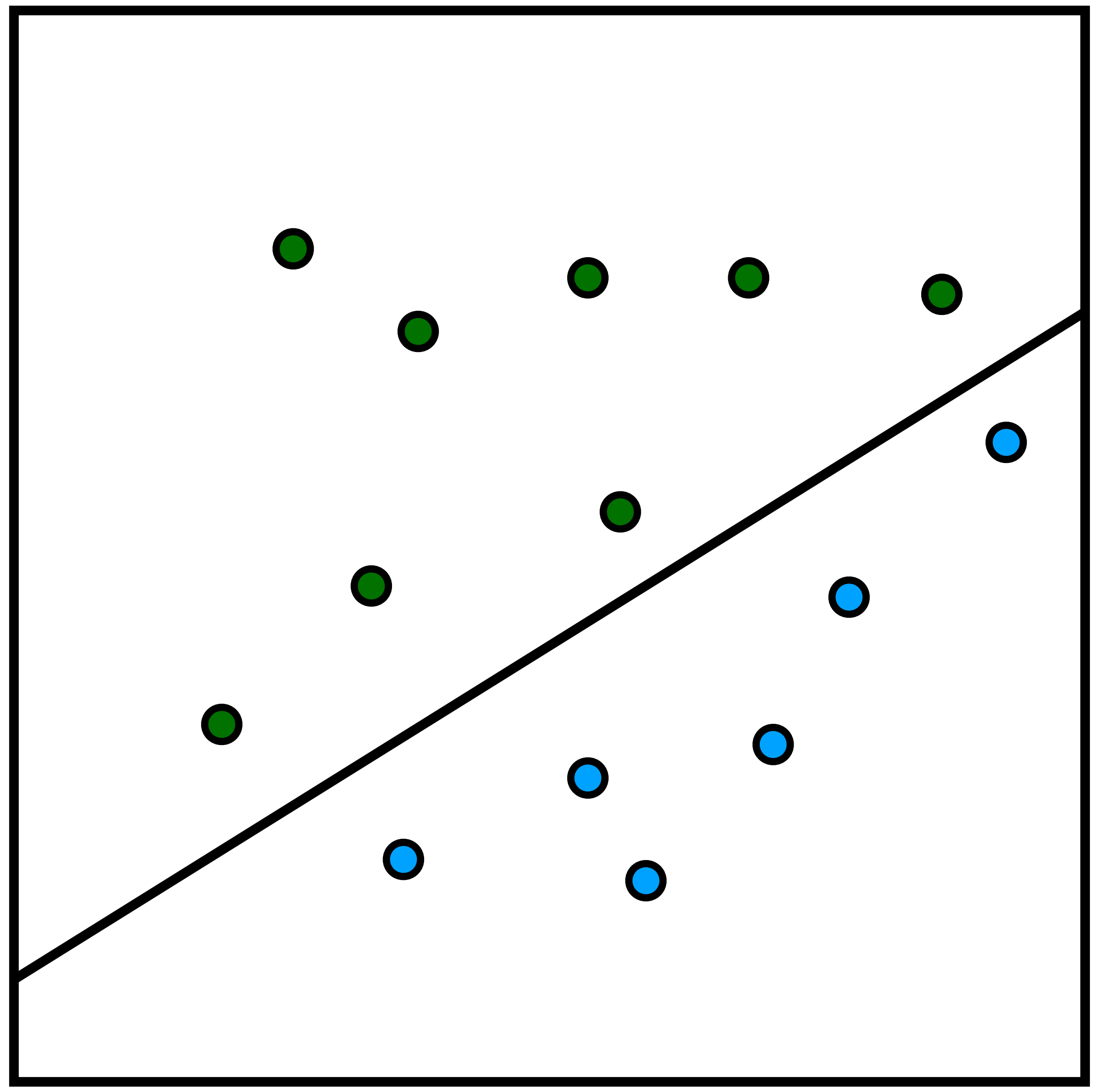


Linear Classification



$$\mathbf{y}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + \mathbf{w}_0$$

Linear Classification



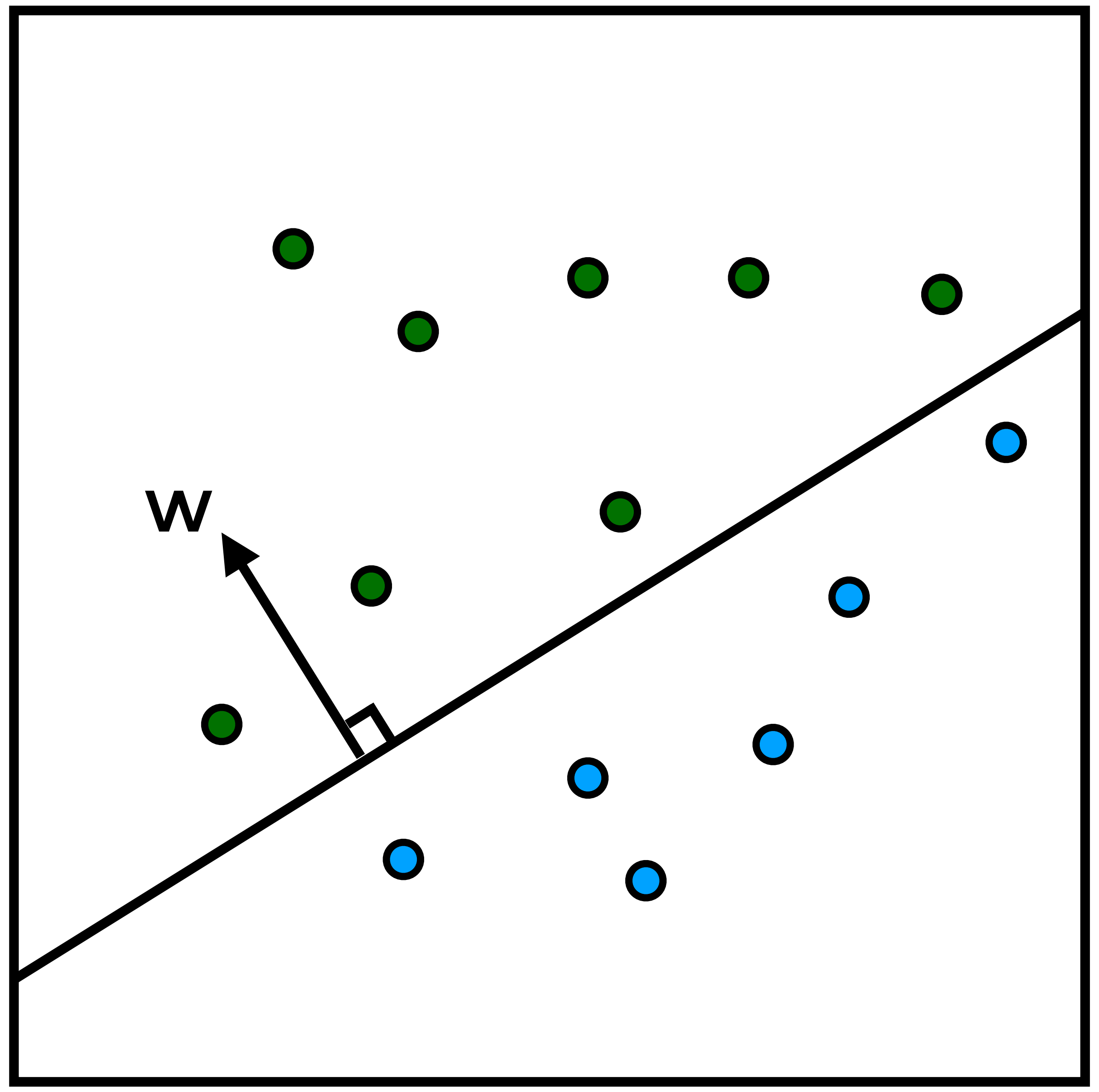
$$\mathbf{y}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + \mathbf{w}_0$$

Decision

$$(\mathbf{w}^\top \mathbf{x} + \mathbf{w}_0) > 0 \implies \text{green}$$

$$(\mathbf{w}^\top \mathbf{x} + \mathbf{w}_0) < 0 \implies \text{blue}$$

Linear Classification



decision boundary: $y(x) = 0$

take two points on the boundary: x_a, x_b

then: $w^\top x_a + w_0 = w^\top x_b + w_0$

$\implies w^\top (x_a - x_b) = 0$

$\implies$ w is perpendicular to the decision boundary

w represents the orientation of the decision boundary

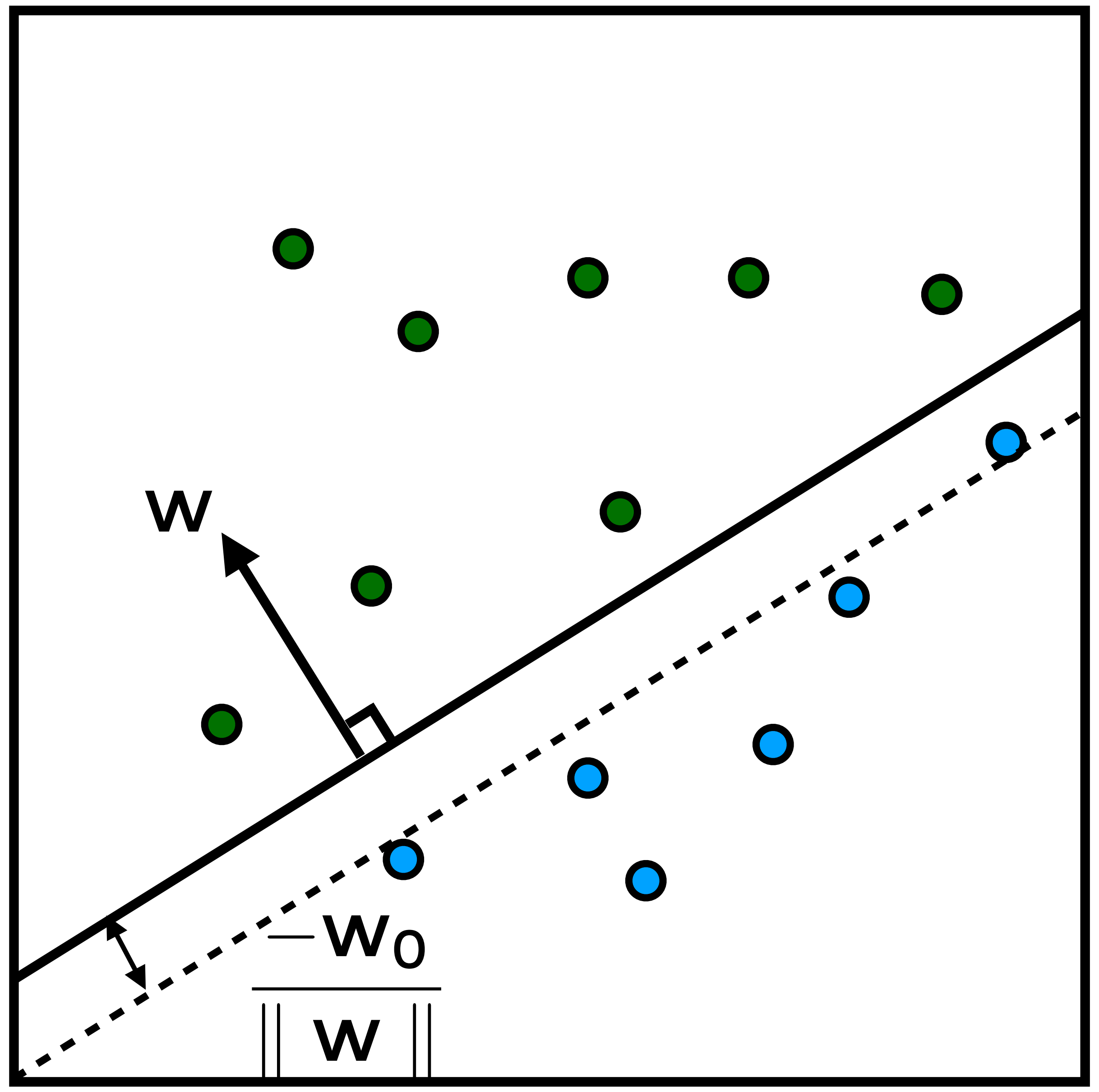
$$y(x) = w^\top x + w_0$$

Decision

$$(w^\top x + w_0) > 0 \implies$$

$$(w^\top x + w_0) < 0 \implies$$

Linear Classification



$$y(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + w_0$$

Decision

$$(\mathbf{w}^\top \mathbf{x} + w_0) > 0 \implies \text{green}$$
$$(\mathbf{w}^\top \mathbf{x} + w_0) < 0 \implies \text{blue}$$

w_0 is a scalar

you can think of it like an intercept

take $\mathbf{x}'$ as the closest point on the decision boundary to the origin

$\mathbf{x}' = \beta \mathbf{w}$

$$\implies y(\mathbf{x}') = \mathbf{w}^\top \mathbf{x}' + w_0$$
$$\implies y(\mathbf{x}') = \mathbf{w}^\top (\beta \mathbf{w}) + w_0$$
$$\implies 0 = \beta \| \mathbf{w} \|^2 + w_0$$
$$\implies \beta = \frac{-w_0}{\| \mathbf{w} \|^2}$$

Then you know that the distance from the origin to $\mathbf{x}'$ is:

$$\| \mathbf{x}' \| = \| \beta \mathbf{w} \|$$
$$\implies \| \mathbf{x}' \| = \beta \| \mathbf{w} \|$$
$$\implies \| \mathbf{x}' \| = \frac{-w_0}{\| \mathbf{w} \|^2} \| \mathbf{w} \|$$
$$\implies \| \mathbf{x}' \| = \frac{-w_0}{\| \mathbf{w} \|}$$

Support Vector Machine (SVM)

Support Vector Machine (SVM)

- **In the previous slide the estimated decision boundary may be affected by hyper parameters (e.g., the order of the dataset, how the parameters were initialized).**

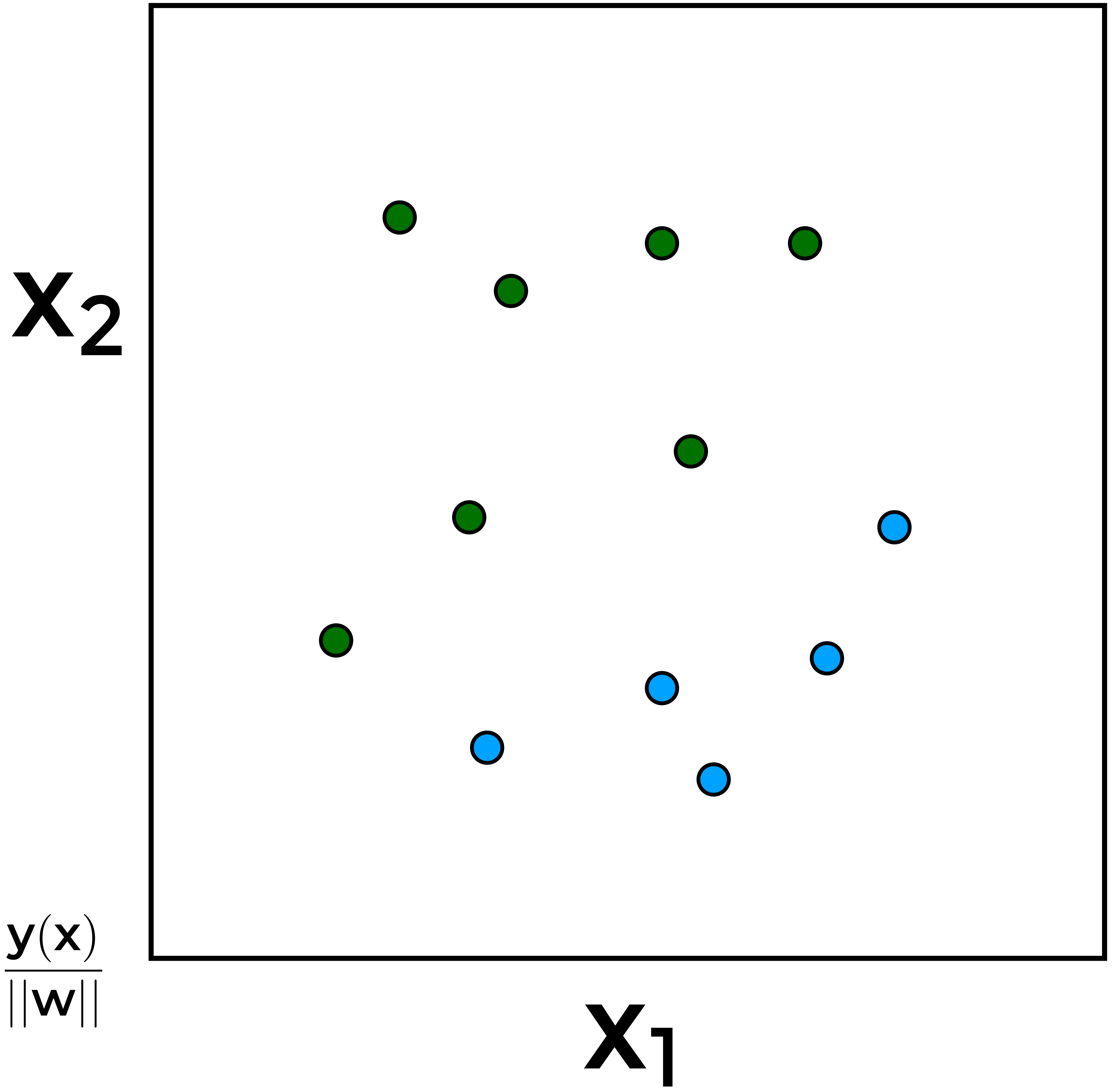
Support Vector Machine (SVM)

- In the previous slide the estimated decision boundary may be affected by hyper parameters (e.g., the order of the dataset, how the parameters were initialized).

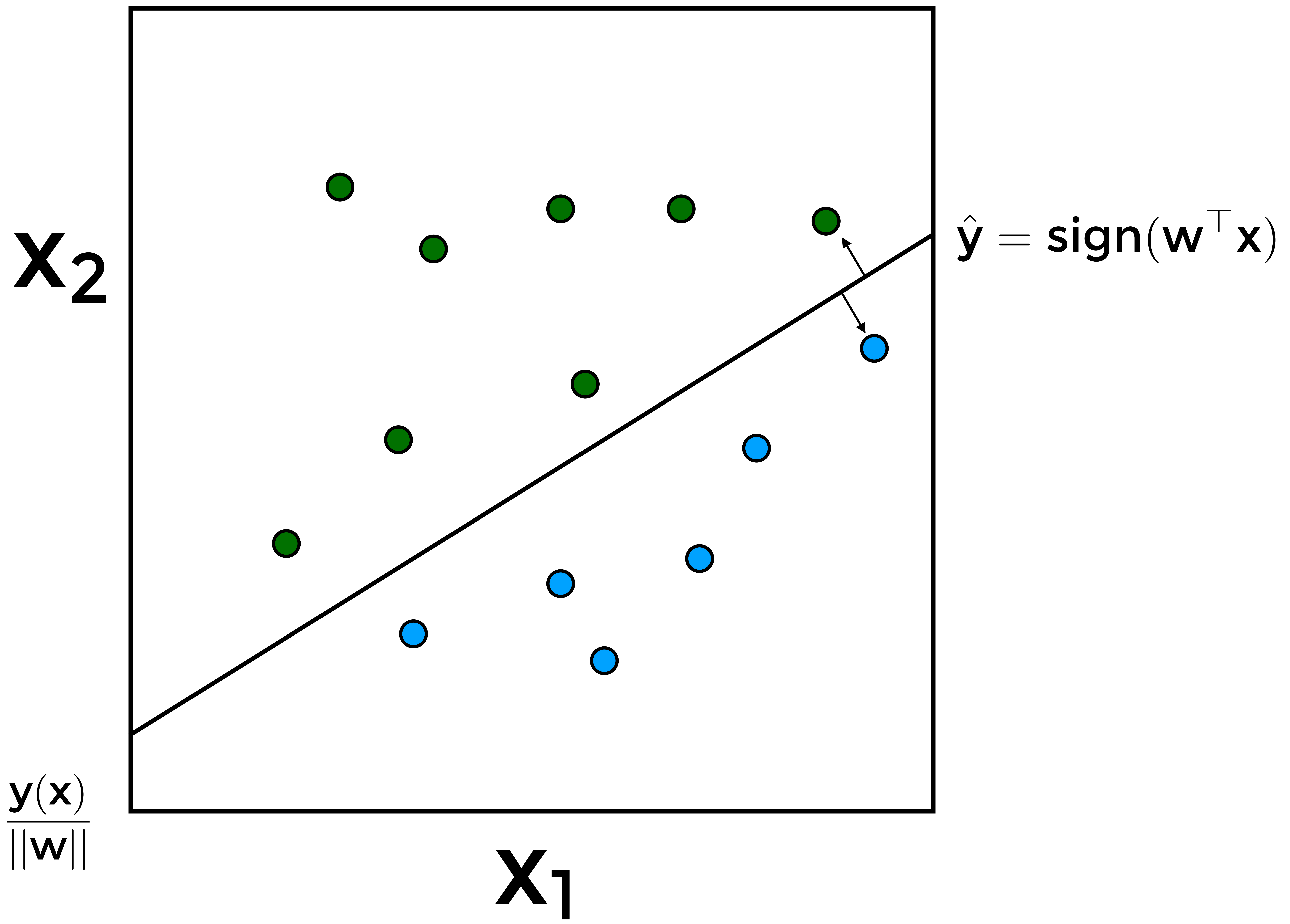
- SVMs aim at finding the decision boundary that maximize the margin

Support Vector Machine (SVM)

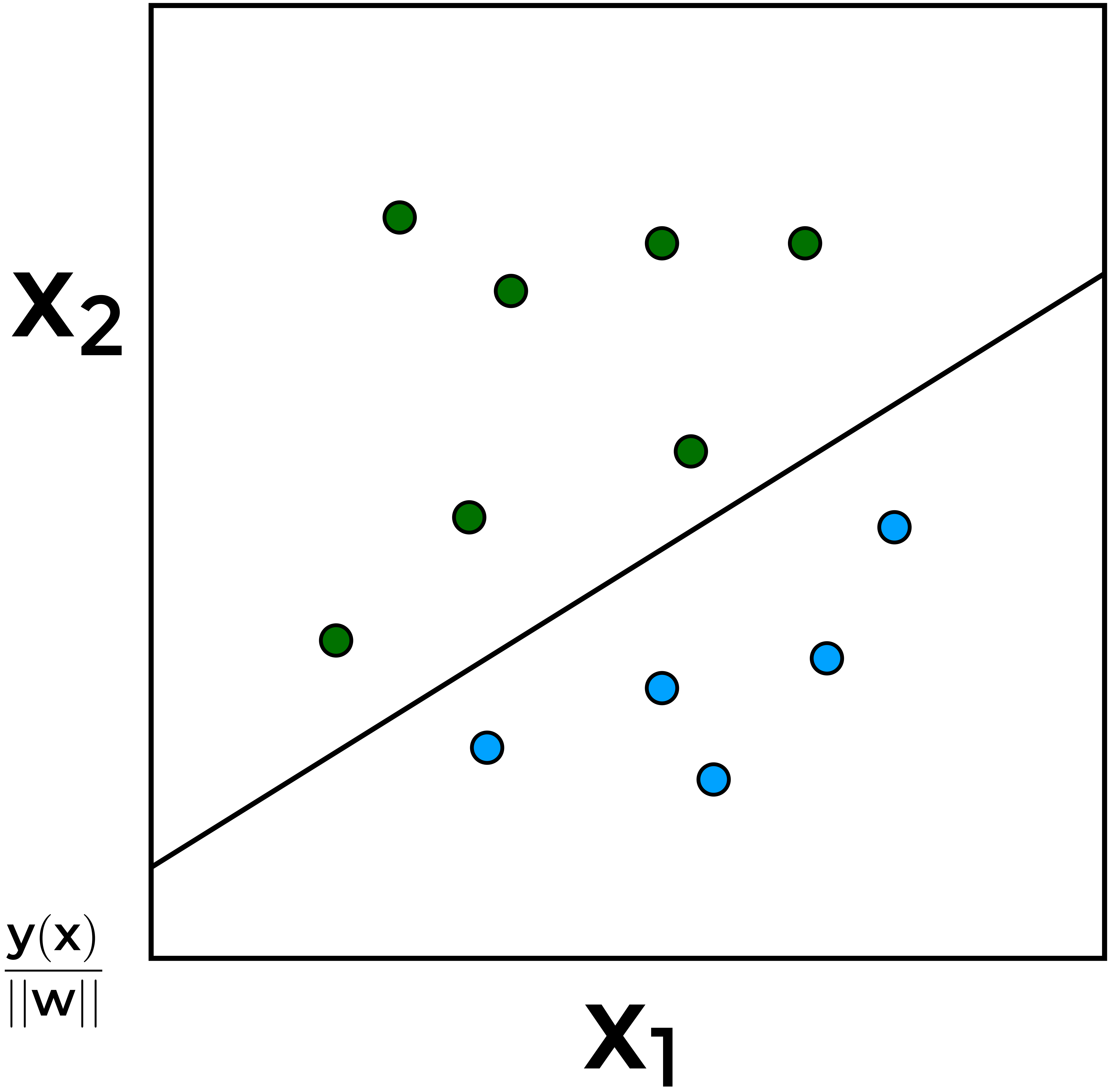
- In the previous slide the estimated decision boundary may be affected by hyper parameters (e.g., the order of the dataset, how the parameters were initialized).
- SVMs aim at finding the decision boundary that maximize the margin
- Popular and powerful approach
  - Comes with theoretical guarantees
  - Results in a convex optimization
  - Ideas extended to structured outputs



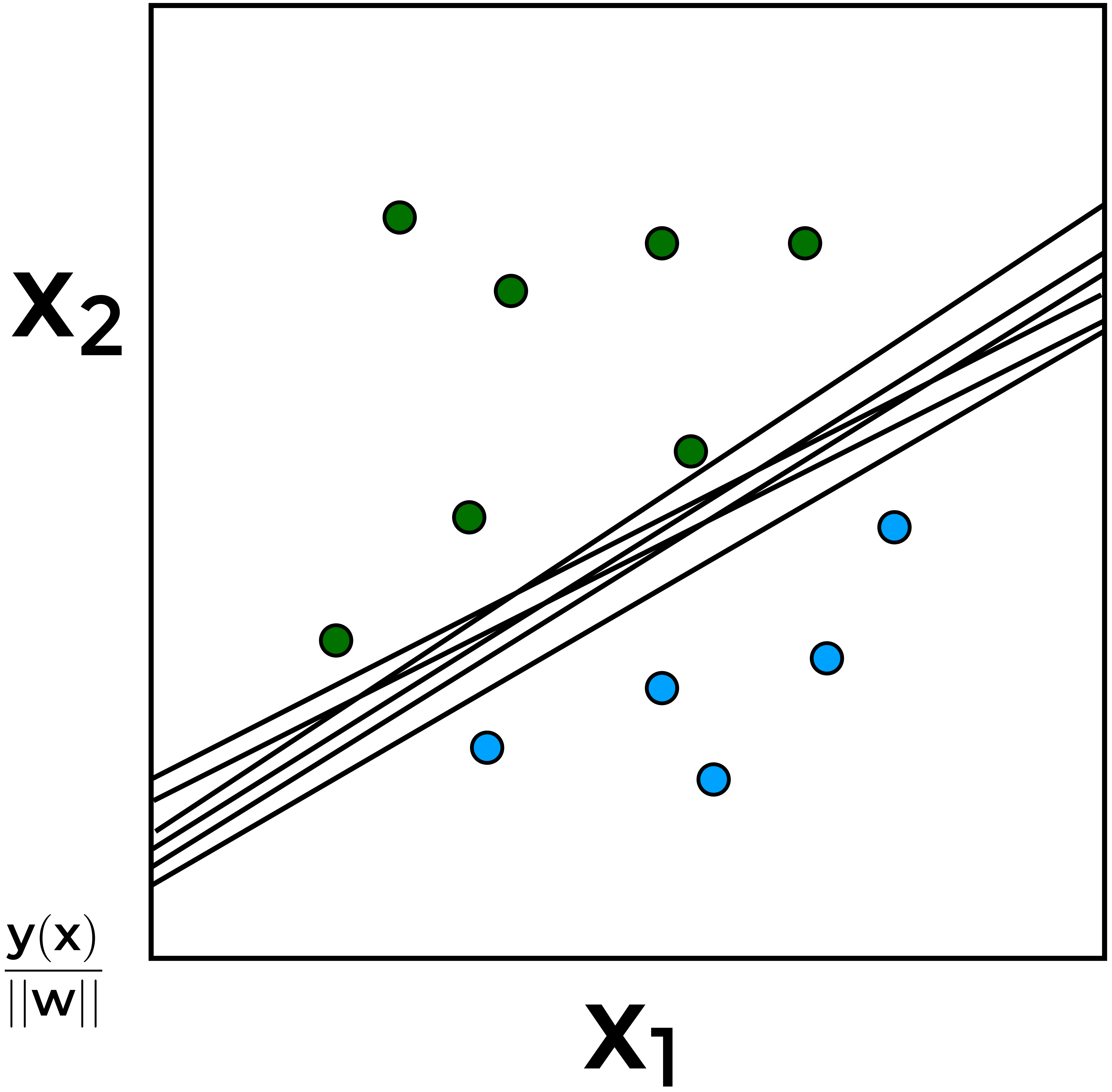
$$\frac{y(x)}{||w||}$$



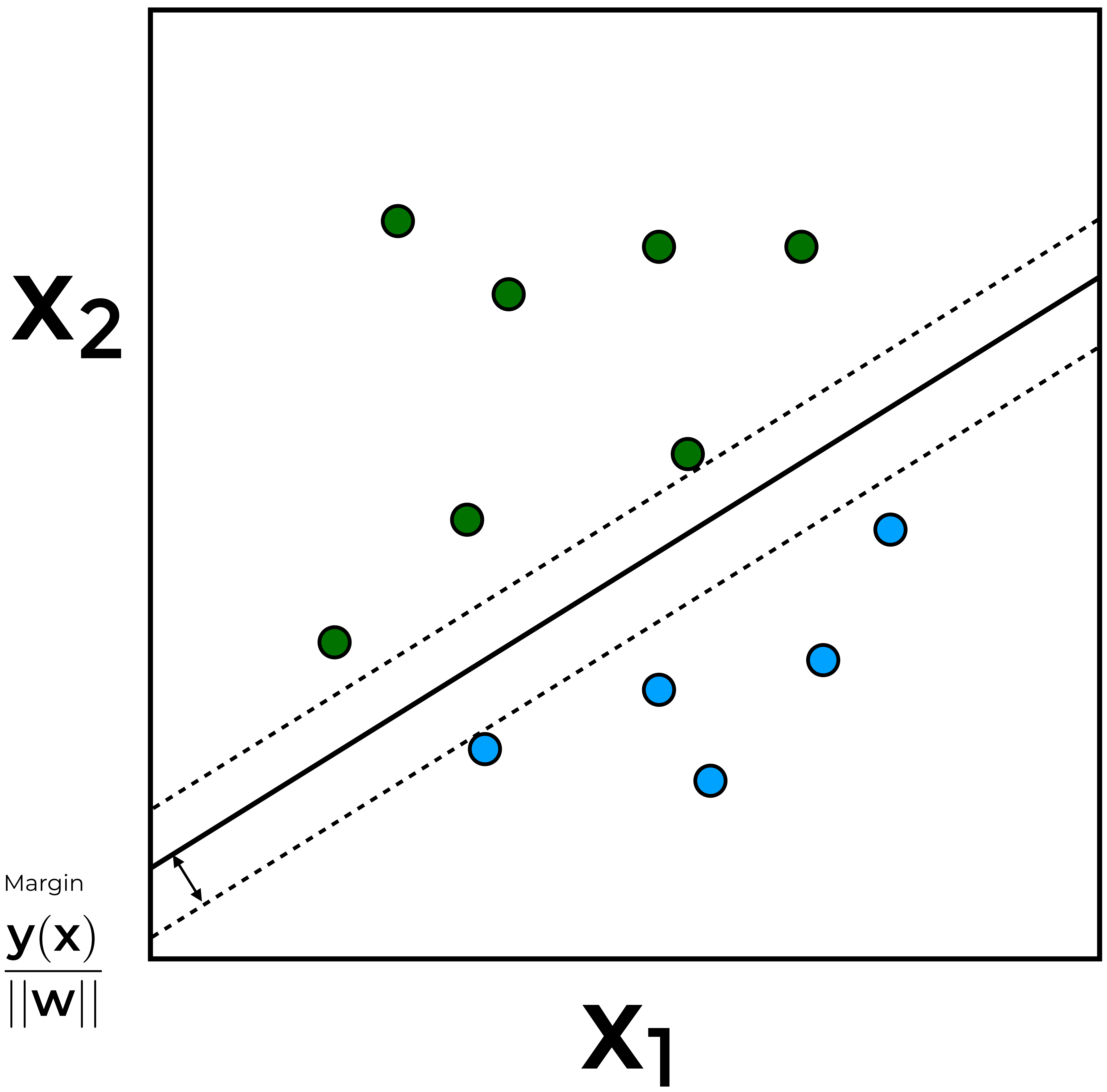
$\mathbf{x}_2$
$\hat{y} = \text{sign}(\mathbf{w}^\top \mathbf{x})$
$\frac{y(\mathbf{x})}{||\mathbf{w}||}$
$\mathbf{x}_1$



x_2
$\frac{y(x)}{||w||}$
x_1

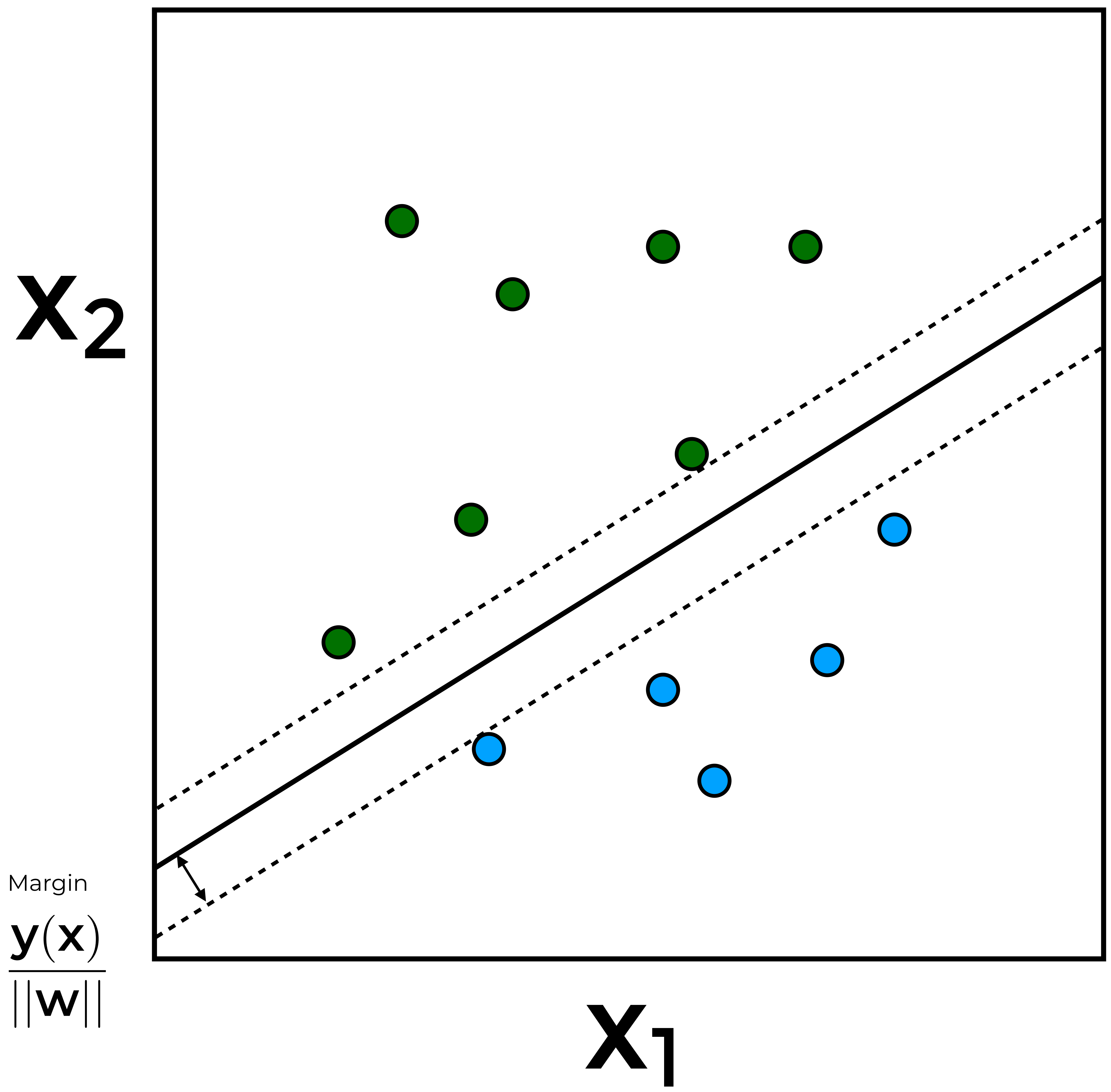


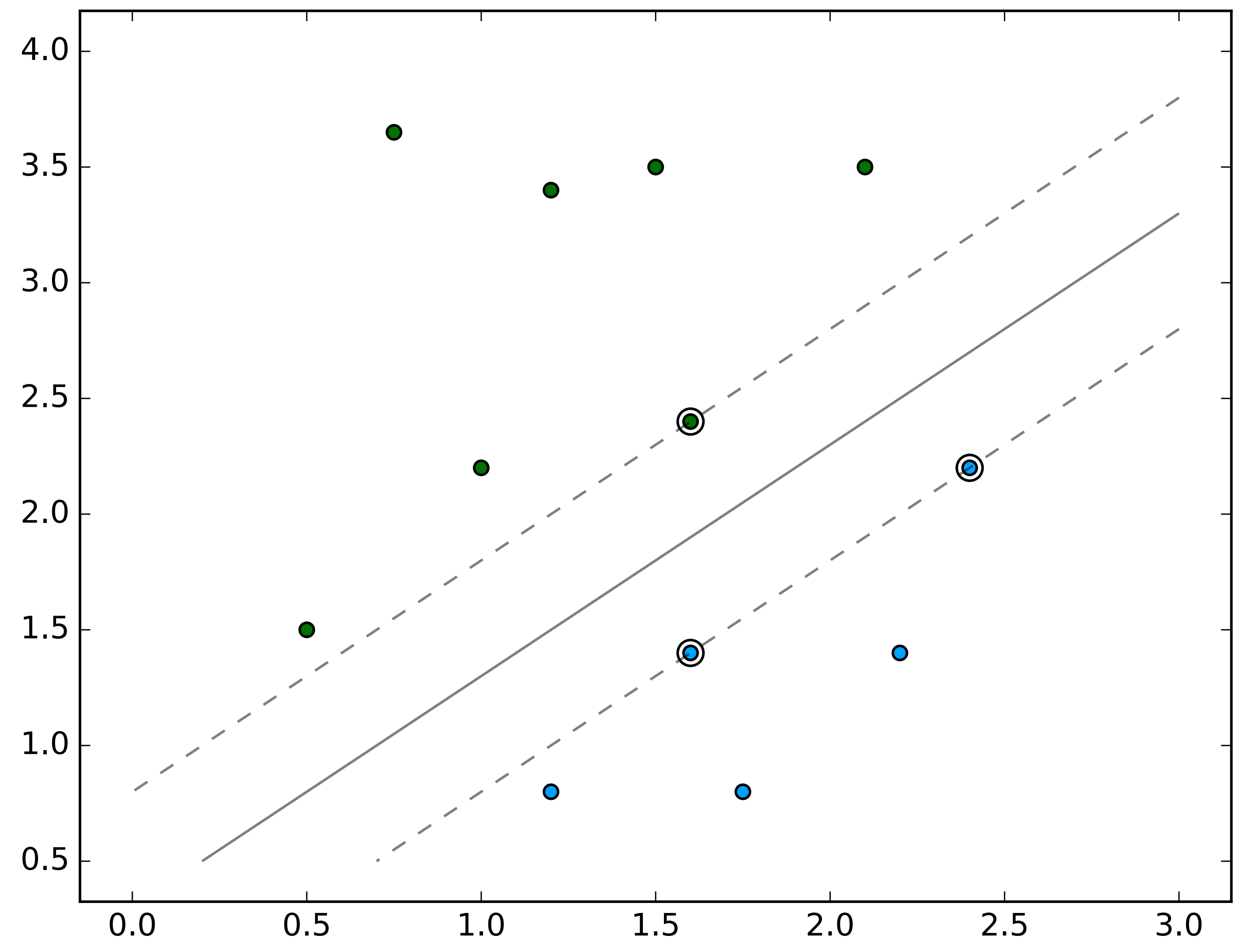
x_2
x_1
$\frac{y(x)}{||w||}$



x_2
x_1
Margin
$\frac{\mathbf{y}(\mathbf{x})}{||\mathbf{w}||}$

The objective is to find the separating boundary that maximizes the margin






```python
# Scikit-learn library
clf = svm.SVC(kernel='linear', C=1000)
clf.fit(X, y)
```

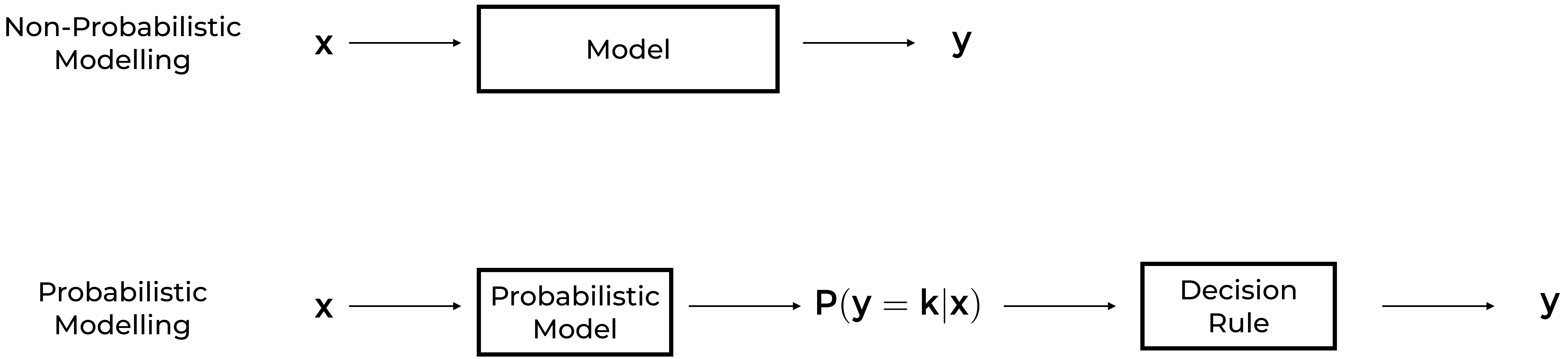
[Using: http://scikit-learn.org/stable/auto_examples/svm/plot_separating_hyperplane.html]

Probabilistic Models
for Classification

Decision and Inference

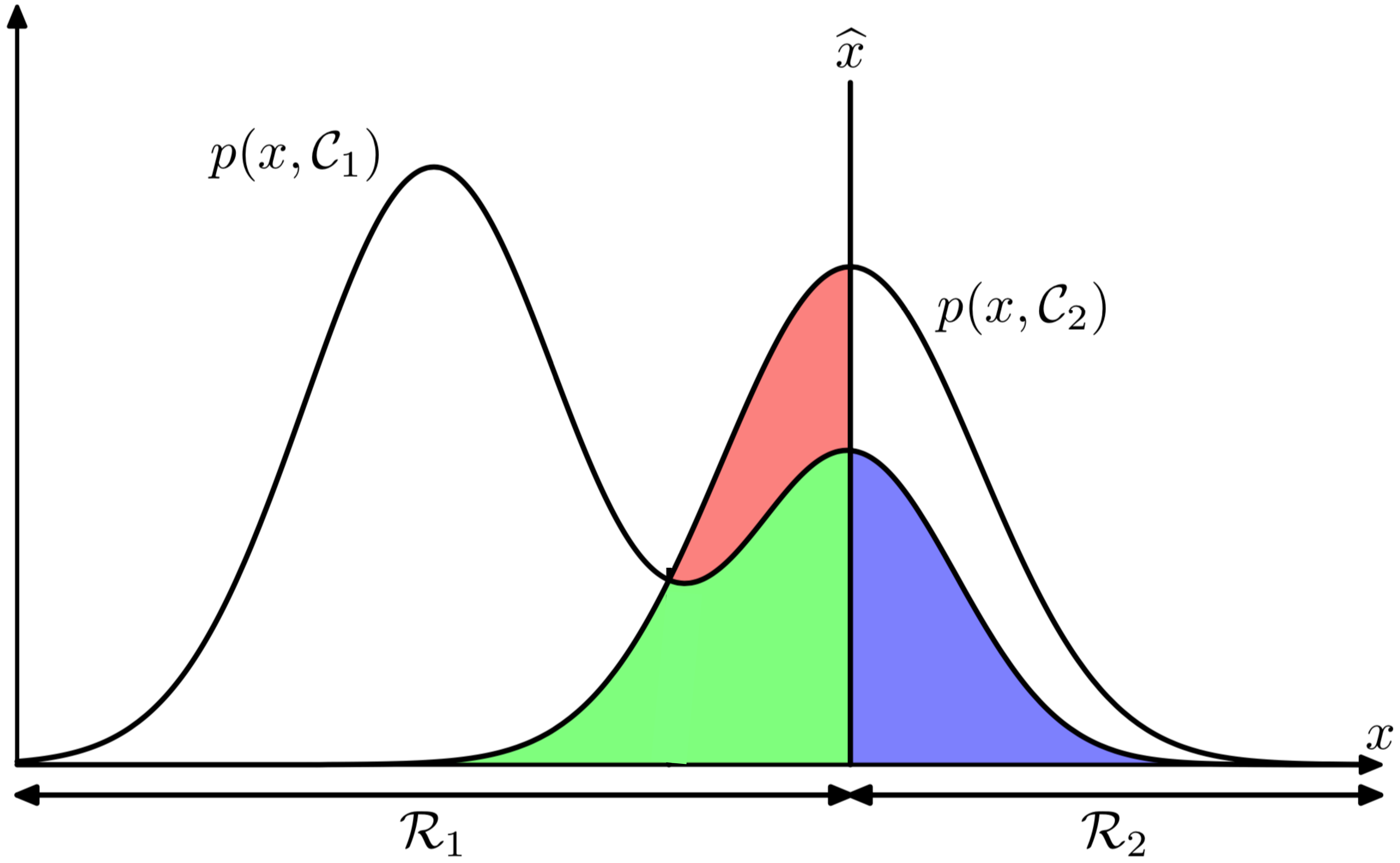
- Classification models provide a class label given a datum
- Probabilistic classification models more clearly divide the problem into two sub-tasks:

  1. Inferring
  2. Making a decision based on the inference results


Non-Probabilistic
Modelling
x
Model
y
Probabilistic
Modelling
x
Probabilistic
Model
$P(y = k|x)$
Decision
Rule
y

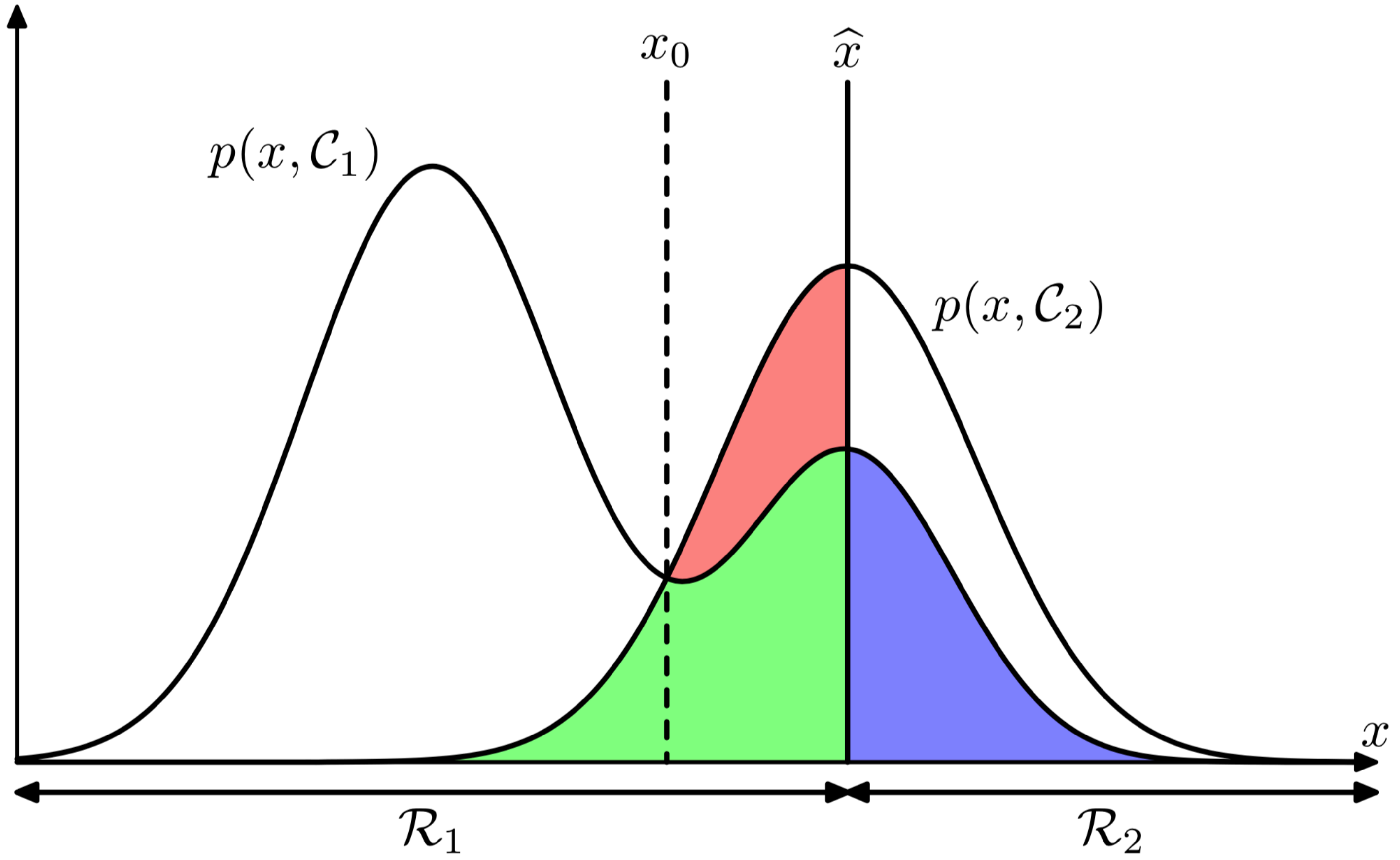
Decision Theory (1 slide)

$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\, \mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\, \mathrm{d}\mathbf{x}. \end{aligned}$$



Decision Theory (1 slide)

$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\,\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\,\mathrm{d}\mathbf{x}. \end{aligned}$$



Extra flexibility

- Separating inference from decision can be useful:
  - Examine (predictive) uncertainty
  - Minimize risk
    - Cost of false pos. differs from cost of false neg.
  - Combine models
  - Compensate for class imbalance

Probabilistic models

1. **Model the conditional directly:**

$$\mathsf{P}(\mathbf{y} = \mathbf{k} | \mathbf{x})$$

2. **Model the joint (or the prior and the class conditionals):**

$$\underbrace{\mathsf{P}(\mathbf{y} = \mathbf{k} | \mathbf{x})}_{\text{posterior}} \propto \underbrace{\mathsf{P}(\mathbf{y} = \mathbf{k}, \mathbf{x})}_{\text{joint}}$$

$\boxed{\text{Bayes' Theorem}}$

$$= \underbrace{\mathsf{P}(\mathbf{x} \mid \mathbf{y} = \mathbf{k})}_{\text{class conditional densities}} \underbrace{\mathsf{P}(\mathbf{y} = \mathbf{k})}_{\text{class prior}}$$

- In the next few slides we will build models from the ground up

- We will show how simple modeling decisions lead to known models

- **In most cases we will parametrize the distributions, e.g.:**

$$\underbrace{\mathsf{P}(\mathbf{y} = \mathbf{k}|\mathbf{x})}_{\text{posterior}} \propto \underbrace{\mathsf{P}(\mathbf{y} = \mathbf{k}, \mathbf{x})}_{\text{joint}}$$
$$= \underbrace{\mathsf{P}(\mathbf{x} \mid \mathbf{y} = \mathbf{k})}_{\text{class conditional densities}} \underbrace{\mathsf{P}(\mathbf{y} = \mathbf{k})}_{\text{class prior}}$$

$$\underbrace{\mathsf{P}(\mathbf{x} \mid \mathbf{y} = \mathbf{k}, \theta)}_{\text{class conditional density}}$$

- **Assume that x is a vector of dimensionality M**

$$\mathbf{x} = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_M \end{bmatrix}$$

Building a model

- Assume that dimensions of x are independent

$$P(x \mid y = k) = \prod_{j=1}^{M} P(x_j \mid y = k, \theta_{jk})$$

  - the problem is then to model each conditional (there are M of them)

  - this model is known as a Naive Bayes classifier

Building a model

- If x is binary: $x_j \in \{0, 1\} \;\; \forall j$.

$$P(x_j \mid y = k, \theta_{jk}) = \text{Bernoulli}(x_j \mid p_{jk}) \qquad\qquad \theta_{jk} := p_{jk}$$

- If x is continuous:

$$P(x_j \mid y = k, \theta_{jk}) = \mathcal{N}(x_j \mid \mu_{jk}, \sigma^2_{jk}) \qquad\qquad \theta_{jk} := \{\mu_{jk}, \sigma^2_{jk}\}$$

- If x is “mix” we can use a different distribution for each dimension

Estimating the parameters (e.g., θ)

- What is our performance measure?
  - Turn the estimation problem into an optimization problem

1. Maximum likelihood estimate (MLE)

2. Maximum a posterior (MAP)

3. Full posterior

Maximum Likelihood (MLE)

Likelihood: $\mathsf{P}(\mathbf{x}, \mathbf{y} \mid \theta)$

$$= \mathsf{P}(\mathbf{x} \mid \mathbf{y}, \beta)\mathsf{P}(\mathbf{y} \mid \boldsymbol{\pi}) \quad \theta = \{\beta, \boldsymbol{\pi}\}$$

$$= \prod_{j}^{M} \mathsf{P}(\mathbf{x}_j \mid \mathbf{y}, \beta)\mathsf{P}(\mathbf{y} \mid \boldsymbol{\pi})$$

- **Parametrize both distributions according to the data type**
  - **E.g., a multinomial for y and a Bernoulli for binary x.**
- **Solve the following optimization problem:**

$$\hat{\theta} = \arg\max_{\theta} \mathsf{P}(\mathbf{x}, \mathbf{y} \mid \theta)$$

- For a binary X and a categorical Y

$$\text{log-likelihood} = \log\left(\prod_{j=1}^{M} P(x_j \mid y, p) P(y \mid \boldsymbol{\pi})\right)$$
$$= \log\left(\prod_{j=1}^{M}\prod_{k=1}^{K} \text{Bernoulli}(x_j \mid p_{jk}) \text{Categorical}(y = k \mid \boldsymbol{\pi}_k)\right)$$
$$= \sum_{j=1}^{M}\sum_{k=1}^{K} P(x_j \mid p_{jk}) + \sum_{k=1}^{K} N_k \log \boldsymbol{\pi}_k$$

- MLE solutions:

$$\hat{p}_{jk} = \frac{N_{jk}}{N_k}$$
$$\hat{\pi}_k = \frac{N_k}{N}$$

N : total number of instances

N_k : number of instances where $y = k$

N_{jk} : number of instances where $y = k$ and $x_j = 1$

Making predictions

- You can then use the MLE estimates for predictions

$\boxed{\text{Bayes' Theorem}}$

$$\underbrace{P(y = k|x)}_{\text{posterior}} \propto \underbrace{P(y = k, x)}_{\text{joint}}$$

$$= \underbrace{P(x \mid y = k)}_{\text{class conditional densities}} \underbrace{P(y = k)}_{\text{class prior}}$$

Making predictions

- You can then use the MLE estimates for predictions

$\boxed{\text{Bayes' Theorem}}$

$$\underbrace{P(y = k|x)}_{\text{posterior}} \propto \underbrace{P(y = k, x)}_{\text{joint}}$$

$$= \underbrace{P(x \mid y = k)}_{\text{class conditional densities}} \underbrace{P(y = k)}_{\text{class prior}}$$

MLE estimates: $\hat{p}_{jk} = \frac{N_{jk}}{N_k}$ $\hat{\pi}_k = \frac{N_k}{N}$

$$\hat{p}_{jk} = \frac{N_{jk}}{N_k}$$

$$\hat{\pi}_k = \frac{N_k}{N}$$

- The MLE estimate relies solely on the training set
  - It provides the best fit to the observed data given the model
  - It can overfit.

Maximum a posterior (MAP)

- As a fix we can model the parameters are R.V.
  - This allows us to encode prior knowledge
    - e.g., all classes have some non-zero probability
- Compared to MLE the MAP procedure takes into account this prior

$$\mathsf{P}(\boldsymbol{\pi}) = \mathsf{Dirichlet}(\alpha_1, \ldots, \alpha_K)$$

$$\hat{\boldsymbol{\pi}}_k = \frac{N_k + \alpha_k}{N + \sum_{k'} \mathrm{alpha}_{k'}}$$

```
>>> import numpy as np
>>> X = np.random.randint(2, size=(6, 100))
>>> Y = np.array([1, 2, 3, 4, 4, 5])
>>> from sklearn.naive_bayes import BernoulliNB
>>> clf = BernoulliNB()
>>> clf.fit(X, Y)
BernoulliNB(alpha=1.0, binarize=0.0,
class_prior=None, fit_prior=True)
>>> print(clf.predict(X[2:3]))
[3]
```

http://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.BernoulliNB.html#sklearn.naive_bayes.BernoulliNB

Complete example

Use Naive Bayes to train a document classifier

- A model that predicts a document's topic (class)
- Document will be encoded using Bag-of-Words

- **Documents are email messages sent to a newsgroup**

```
From: bcash@crchh410.NoSubdomain.NoDomain (Brian Cash)
Subject: Re: free moral agency
Nntp-Posting-Host: crchh410
Organization: BNR, Inc.
Lines: 17

In article <735295730.25282@minster.york.ac.uk>, cjhs@minster.york.ac.uk writes:
|> : Are you saying that their was a physical Adam and Eve, and that all
|> : humans are direct decendents of only these two human beings.?  Then who
|> : were Cain and Able's wives?  Couldn't be their sisters, because A&E
|> : didn't have daughters.  Were they non-humans?
|>
|> Genesis 5:4
|>
|> and the days of Adam after he begat Seth were eight hundred years, and
|> he begat sons and daughters:
|>
|> Felicitations -- Chris Ho-Stuart

Yeah, but these were not the wives.  The wives came from Nod, apparently
a land being developed by another set of gods.

Brian /-|-\
```

Document 40

- Classes correspond to newsgroup topic
  - Each document belongs to a single class

```
From: bcash@crchh410.NoSubdomain.NoDomain (Brian Cash)
Subject: Re: free moral agency
Nntp-Posting-Host: crchh410
Organization: BNR, Inc.
Lines: 17

In article <735295730.25282@minster.york.ac.uk>, cjhs@minster.york.ac.uk writes:
|> : Are you saying that their was a physical Adam and Eve, and that all
|> : humans are direct decendents of only these two human beings.?  Then who
|> : were Cain and Able's wives?  Couldn't be their sisters, because A&E
|> : didn't have daughters.  Were they non-humans?
|>
|> Genesis 5:4
|>
|> and the days of Adam after he begat Seth were eight hundred years, and
|> he begat sons and daughters:
|>
|> Felicitations -- Chris Ho-Stuart

Yeah, but these were not the wives.  The wives came from Nod, apparently
a land being developed by another set of gods.

Brian /-|-\
```

Document 40

20 classes

comp.graphics
comp.os.ms-windows.misc
comp.sys.ibm.pc.hardware
comp.sys.mac.hardware
comp.windows.x

rec.autos
rec.motorcycles
rec.sport.baseball
rec.sport.hockey

sci.crypt
sci.electronics
sci.med
sci.space

misc.forsale

talk.politics.misc
talk.politics.guns
talk.politics.mideast

talk.religion.misc
alt.atheism
soc.religion.christian

- **Bag-of-word encoding encodes the frequency of words (forgets word order, syntax, etc.)**

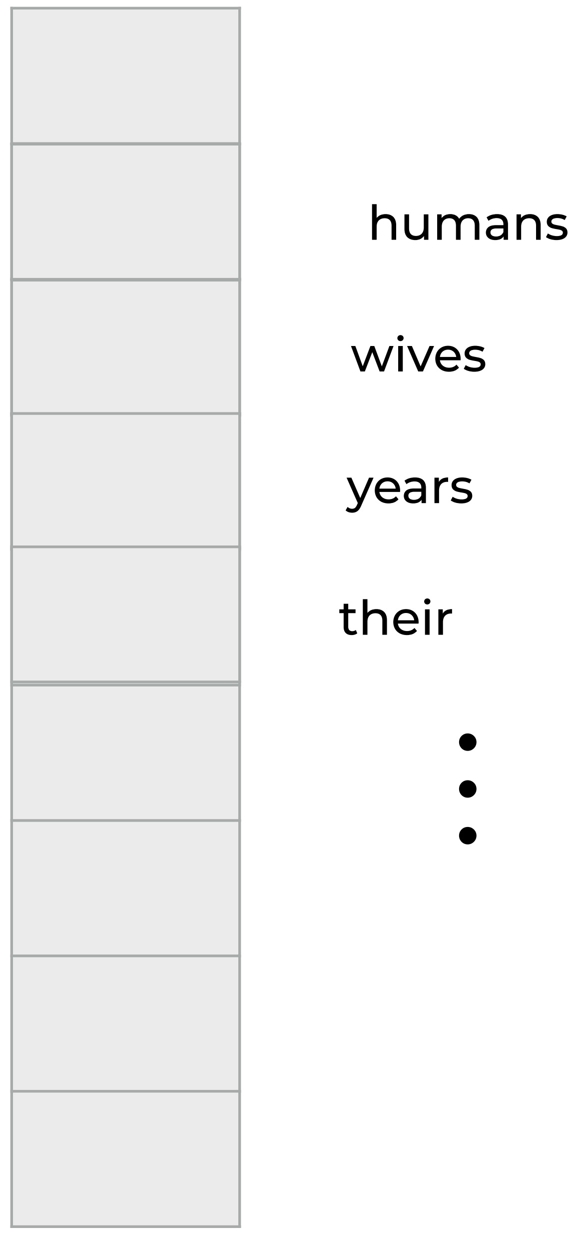
Document 40

```
From: bcash@crchh410.NoSubdomain.NoDomain (Brian Cash)
Subject: Re: free moral agency
Nntp-Posting-Host: crchh410
Organization: BNR, Inc.
Lines: 17

In article <735295730.25282@minster.york.ac.uk>, cjhs@minster.york.ac.uk writes:
|> : Are you saying that their was a physical Adam and Eve, and that all
|> : humans are direct decendents of only these two human beings.?  Then who
|> : were Cain and Able's wives?  Couldn't be their sisters, because A&E
|> : didn't have daughters.  Were they non-humans?
|>
|> Genesis 5:4
|>
|> and the days of Adam after he begat Seth were eight hundred years, and
|> he begat sons and daughters:
|>
|> Felicitations -- Chris Ho-Stuart

Yeah, but these were not the wives.  The wives came from Nod, apparently
a land being developed by another set of gods.

Brian /-|-\
```



Vocabulary: 61,168 words

- **Bag-of-word encoding encodes the frequency of words (forgets word order, syntax, etc.)**

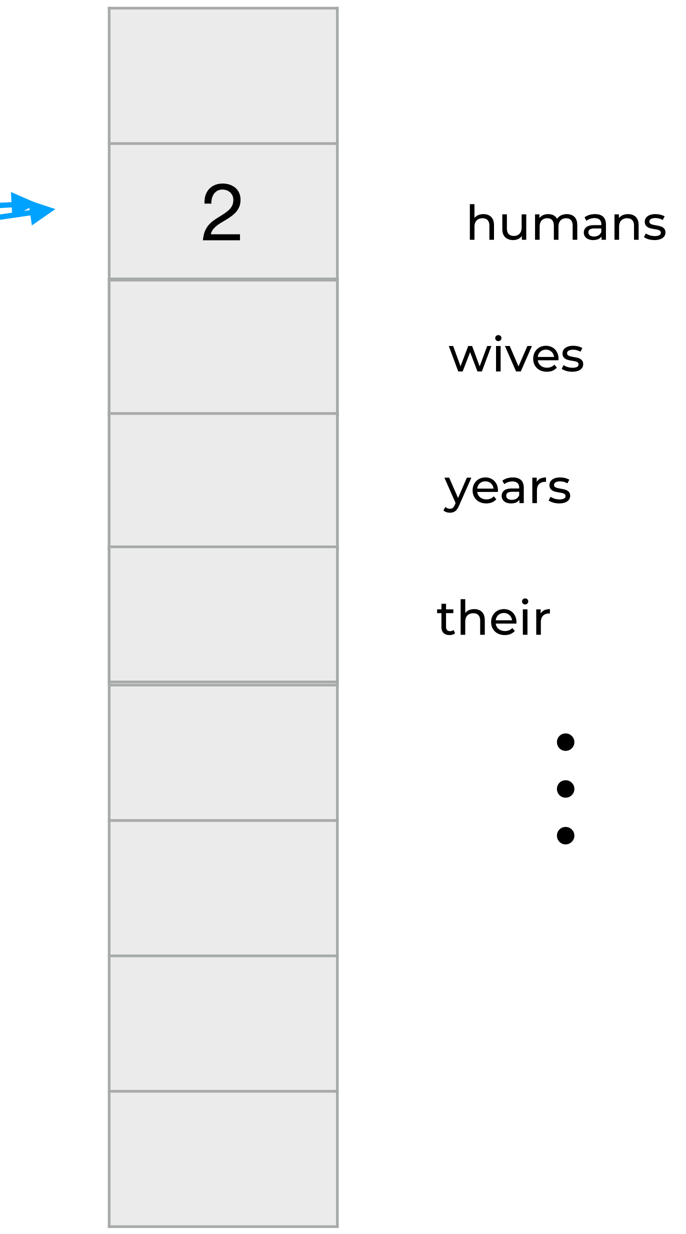
<u>Document 40</u>

```
From: bcash@crchh410.NoSubdomain.NoDomain (Brian Cash)
Subject: Re: free moral agency
Nntp-Posting-Host: crchh410
Organization: BNR, Inc.
Lines: 17

In article <735295730.25282@minster.york.ac.uk>, cjhs@minster.york.ac.uk writes:
|> : Are you saying that their was a physical Adam and Eve, and that all
|> : humans are direct decendents of only these two human beings.?  Then who
|> : were Cain and Able's wives?  Couldn't be their sisters, because A&E
|> : didn't have daughters.  Were they non-humans?
|>
|> Genesis 5:4
|>
|> and the days of Adam after he begat Seth were eight hundred years, and
|> he begat sons and daughters:
|>
|> Felicitations -- Chris Ho-Stuart

Yeah, but these were not the wives.  The wives came from Nod, apparently
a land being developed by another set of gods.

Brian /-|-\
```



Vocabulary: 61,168 words

- **Bag-of-word encoding encodes the frequency of words (forgets word order, syntax, etc.)**

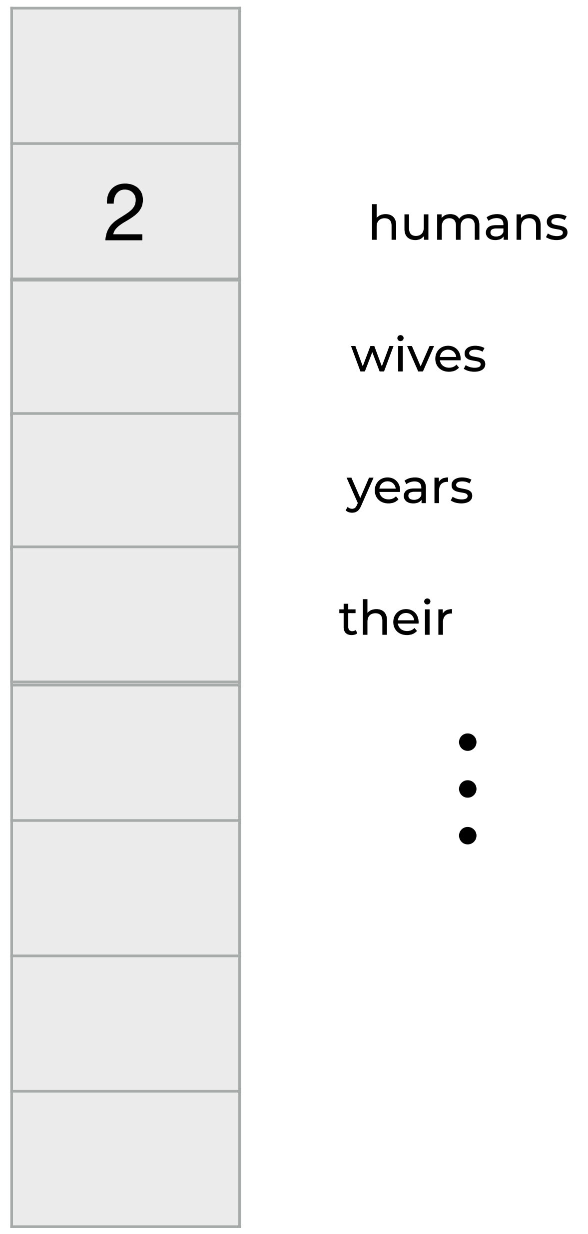
<u>Document 40</u>

```
From: bcash@crchh410.NoSubdomain.NoDomain (Brian Cash)
Subject: Re: free moral agency
Nntp-Posting-Host: crchh410
Organization: BNR, Inc.
Lines: 17

In article <735295730.25282@minster.york.ac.uk>, cjhs@minster.york.ac.uk writes:
|> : Are you saying that their was a physical Adam and Eve, and that all
|> : humans are direct decendents of only these two human beings.?  Then who
|> : were Cain and Able's wives?  Couldn't be their sisters, because A&E
|> : didn't have daughters.  Were they non-humans?
|>
|> Genesis 5:4
|>
|> and the days of Adam after he begat Seth were eight hundred years, and
|> he begat sons and daughters:
|>
|> Felicitations -- Chris Ho-Stuart

Yeah, but these were not the wives.  The wives came from Nod, apparently
a land being developed by another set of gods.

Brian /-|-\
```



Vocabulary: 61,168 words

- **Bag-of-word encoding encodes the frequency of words (forgets word order, syntax, etc.)**

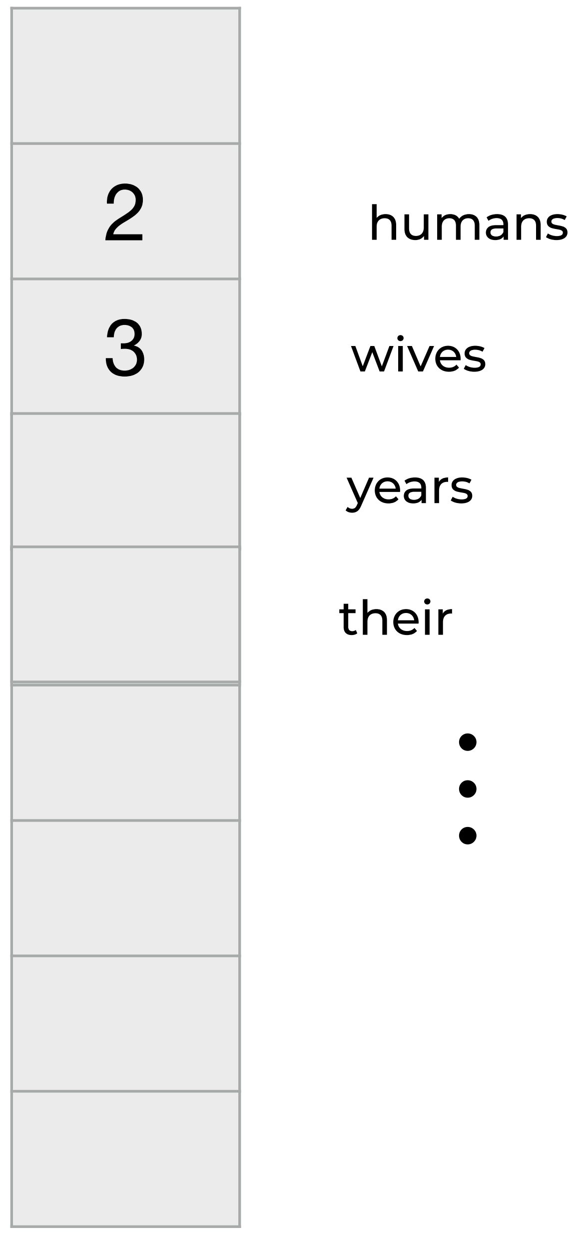
Document 40

```
From: bcash@crchh410.NoSubdomain.NoDomain (Brian Cash)
Subject: Re: free moral agency
Nntp-Posting-Host: crchh410
Organization: BNR, Inc.
Lines: 17

In article <735295730.25282@minster.york.ac.uk>, cjhs@minster.york.ac.uk writes:
|> : Are you saying that their was a physical Adam and Eve, and that all
|> : humans are direct decendents of only these two human beings.?  Then who
|> : were Cain and Able's wives?  Couldn't be their sisters, because A&E
|> : didn't have daughters.  Were they non-humans?
|>
|> Genesis 5:4
|>
|> and the days of Adam after he begat Seth were eight hundred years, and
|> he begat sons and daughters:
|>
|> Felicitations -- Chris Ho-Stuart

Yeah, but these were not the wives.  The wives came from Nod, apparently
a land being developed by another set of gods.

Brian /-|-\
```



Vocabulary: 61,168 words

- **Bag-of-word encoding encodes the frequency of words (forgets word order, syntax, etc.)**

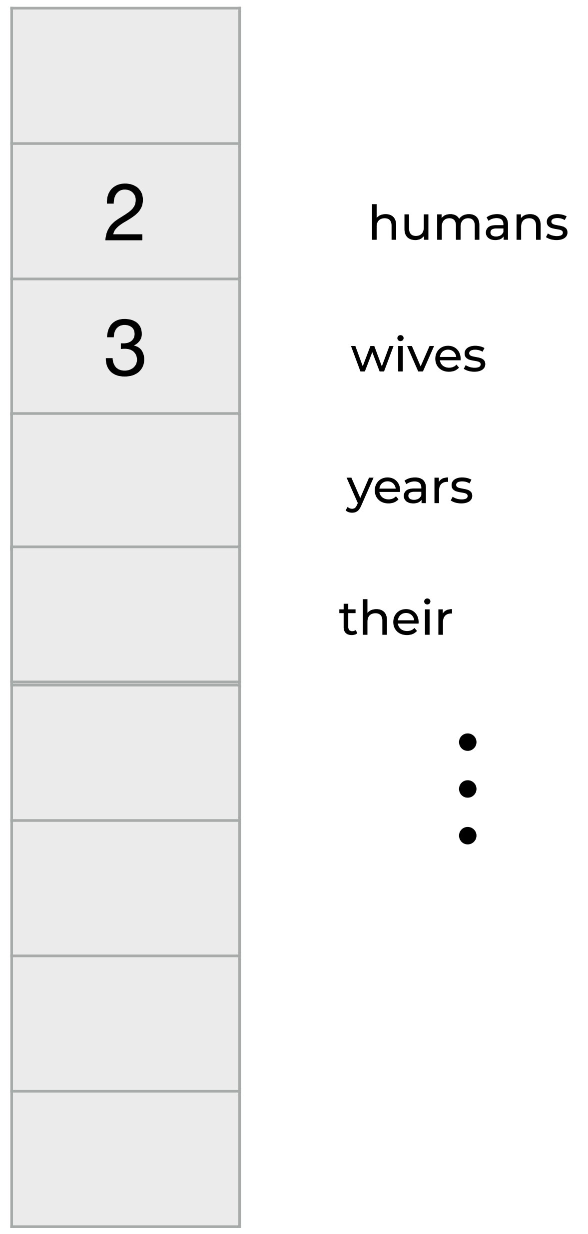
<u>Document 40</u>

```
From: bcash@crchh410.NoSubdomain.NoDomain (Brian Cash)
Subject: Re: free moral agency
Nntp-Posting-Host: crchh410
Organization: BNR, Inc.
Lines: 17

In article <735295730.25282@minster.york.ac.uk>, cjhs@minster.york.ac.uk writes:
|> : Are you saying that their was a physical Adam and Eve, and that all
|> : humans are direct decendents of only these two human beings.?  Then who
|> : were Cain and Able's wives?  Couldn't be their sisters, because A&E
|> : didn't have daughters.  Were they non-humans?
|>
|> Genesis 5:4
|>
|> and the days of Adam after he begat Seth were eight hundred years, and
|> he begat sons and daughters:
|>
|> Felicitations -- Chris Ho-Stuart

Yeah, but these were not the wives.  The wives came from Nod, apparently
a land being developed by another set of gods.

Brian /-|-\
```



Vocabulary: 61,168 words

- **Bag-of-word encoding encodes the frequency of words (forgets word order, syntax, etc.)**

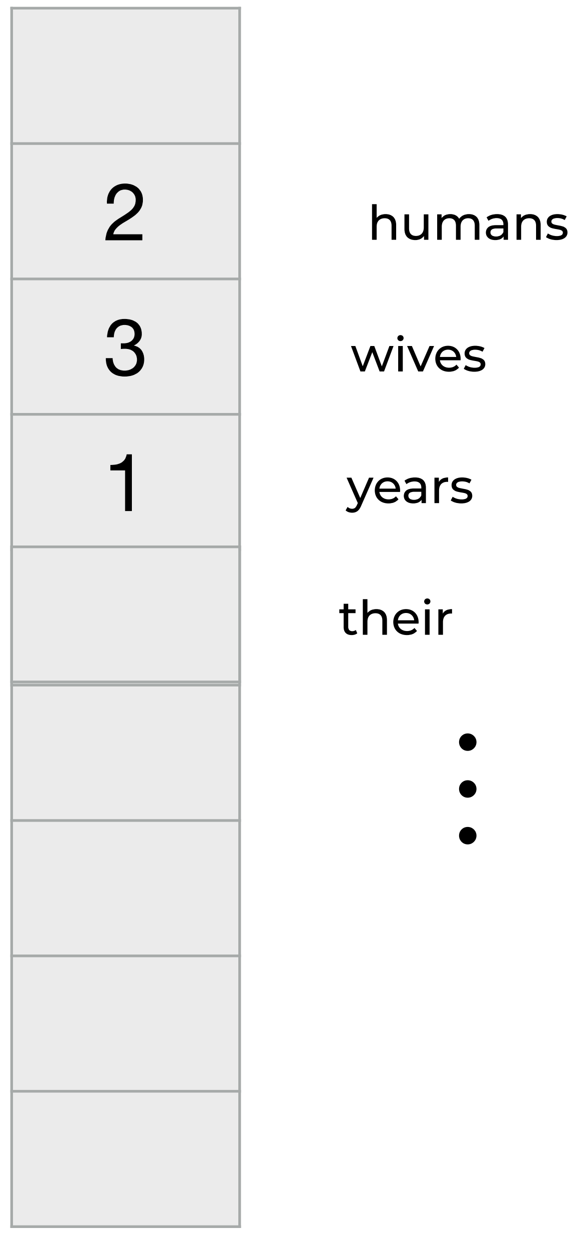
Document 40

```
From: bcash@crchh410.NoSubdomain.NoDomain (Brian Cash)
Subject: Re: free moral agency
Nntp-Posting-Host: crchh410
Organization: BNR, Inc.
Lines: 17

In article <735295730.25282@minster.york.ac.uk>, cjhs@minster.york.ac.uk writes:
|> : Are you saying that their was a physical Adam and Eve, and that all
|> : humans are direct decendents of only these two human beings.?  Then who
|> : were Cain and Able's wives?  Couldn't be their sisters, because A&E
|> : didn't have daughters.  Were they non-humans?
|>
|> Genesis 5:4
|>
|> and the days of Adam after he begat Seth were eight hundred years, and
|> he begat sons and daughters:
|>
|> Felicitations -- Chris Ho-Stuart

Yeah, but these were not the wives.  The wives came from Nod, apparently
a land being developed by another set of gods.

Brian /-|-\
```



Vocabulary: 61,168 words

- **Bag-of-word encoding encodes the frequency of words (forgets word order, syntax, etc.)**

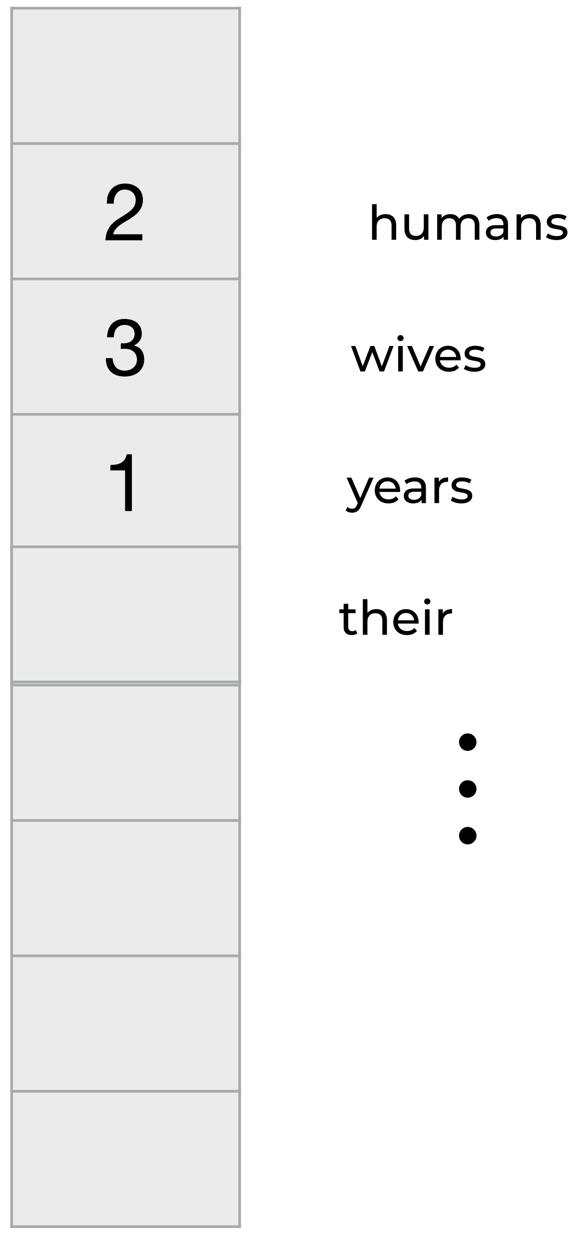
Document 40

```
From: bcash@crchh410.NoSubdomain.NoDomain (Brian Cash)
Subject: Re: free moral agency
Nntp-Posting-Host: crchh410
Organization: BNR, Inc.
Lines: 17

In article <735295730.25282@minster.york.ac.uk>, cjhs@minster.york.ac.uk writes:
|> : Are you saying that their was a physical Adam and Eve, and that all
|> : humans are direct decendents of only these two human beings.?  Then who
|> : were Cain and Able's wives?  Couldn't be their sisters, because A&E
|> : didn't have daughters.  Were they non-humans?
|>
|> Genesis 5:4
|>
|> and the days of Adam after he begat Seth were eight hundred years, and
|> he begat sons and daughters:
|>
|> Felicitations -- Chris Ho-Stuart

Yeah, but these were not the wives.  The wives came from Nod, apparently
a land being developed by another set of gods.

Brian /-|-\
```



Vocabulary: 61,168 words

- **Bag-of-word encoding encodes the frequency of words (forgets word order, syntax, etc.)**

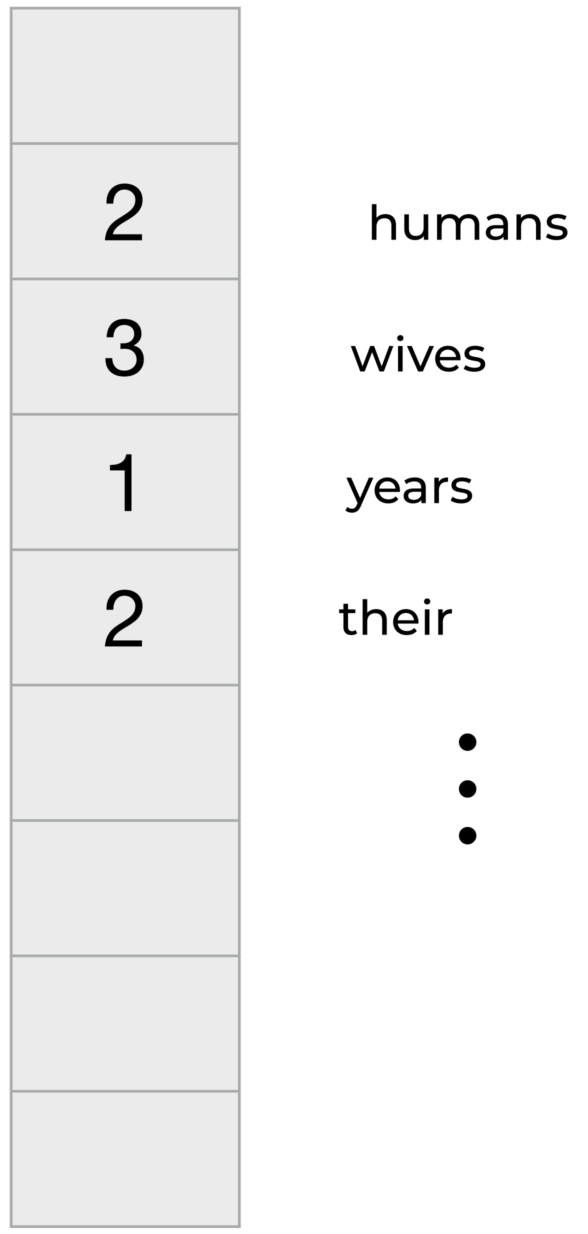
<u>Document 40</u>

```
From: bcash@crchh410.NoSubdomain.NoDomain (Brian Cash)
Subject: Re: free moral agency
Nntp-Posting-Host: crchh410
Organization: BNR, Inc.
Lines: 17

In article <735295730.25282@minster.york.ac.uk>, cjhs@minster.york.ac.uk writes:
|> : Are you saying that their was a physical Adam and Eve, and that all
|> : humans are direct decendents of only these two human beings.?  Then who
|> : were Cain and Able's wives?  Couldn't be their sisters, because A&E
|> : didn't have daughters.  Were they non-humans?
|>
|> Genesis 5:4
|>
|> and the days of Adam after he begat Seth were eight hundred years, and
|> he begat sons and daughters:
|>
|> Felicitations -- Chris Ho-Stuart

Yeah, but these were not the wives.  The wives came from Nod, apparently
a land being developed by another set of gods.

Brian /-|-\
```



Vocabulary: 61,168 words

- **Bag-of-word encoding encodes the frequency of words (forgets word order, syntax, etc.)**

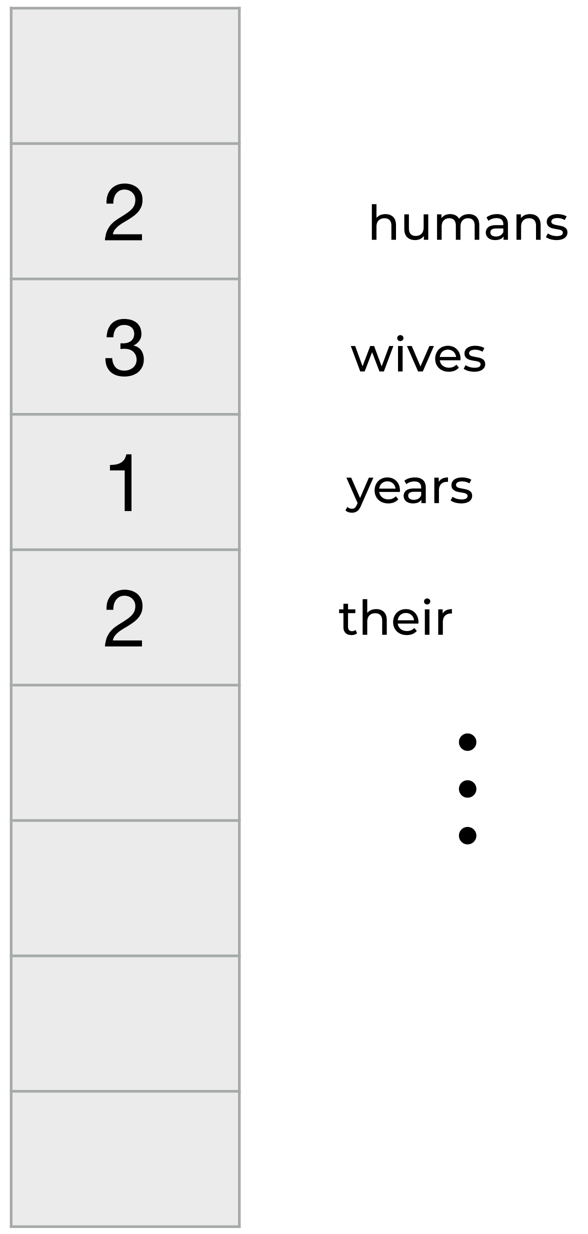
<u>Document 40</u>

```
From: bcash@crchh410.NoSubdomain.NoDomain (Brian Cash)
Subject: Re: free moral agency
Nntp-Posting-Host: crchh410
Organization: BNR, Inc.
Lines: 17

In article <735295730.25282@minster.york.ac.uk>, cjhs@minster.york.ac.uk writes:
|> : Are you saying that their was a physical Adam and Eve, and that all
|> : humans are direct decendents of only these two human beings.?  Then who
|> : were Cain and Able's wives?  Couldn't be their sisters, because A&E
|> : didn't have daughters.  Were they non-humans?
|>
|> Genesis 5:4
|>
|> and the days of Adam after he begat Seth were eight hundred years, and
|> he begat sons and daughters:
|>
|> Felicitations -- Chris Ho-Stuart

Yeah, but these were not the wives.  The wives came from Nod, apparently
a land being developed by another set of gods.

Brian /-|-\
```



Vocabulary: 61,168 words

Data
- 20,000 documents
- 20 classes
- 61,168 vocabulary size

$$\text{Documents} = \begin{bmatrix} 3 & 4 & \cdots & 0 \\ 1 & 0 & \cdots & 9 \\ \vdots & & \ddots & \cdots \\ 0 & 2 & \cdots & 0 \end{bmatrix}_{20,000 \times 61,168}$$

$$\text{Classes} = \begin{bmatrix} 10 \\ 5 \\ \vdots \\ 2 \end{bmatrix}_{20,000 \times 1}$$

Model
- Naive Bayes
- Fit using MAP

$$P(x \mid y = k, p_k) = \text{Multinomial}(x \mid p_k)$$

$$P(y) = \text{Categorical}(\boldsymbol{\pi})$$

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\alpha)$$

Code

https://github.com/lcharlin/80-629/blob/master/inClass/NaiveBayes%2Bexample.ipynb

Beyond Naive

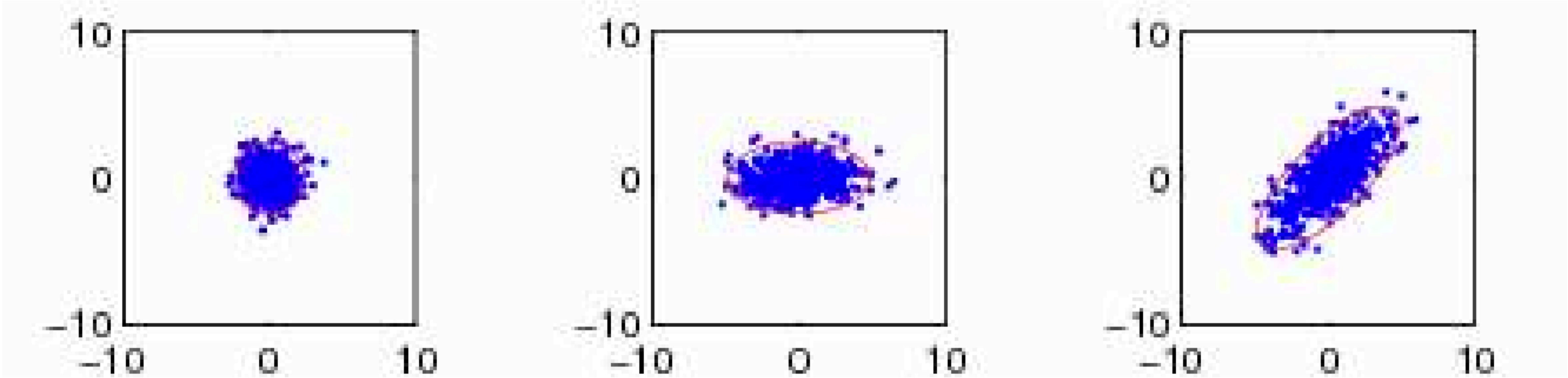
- The assumption behind NB is that features are independent of one another conditioned on the class

$$P(x \mid y = k) = \mathcal{N}(\theta_k, \sigma^2 I)$$

- Unrealistic. e.g., “nasa” and “space”

- There are alternatives specific to continuous X

Quick word on covariance matrices

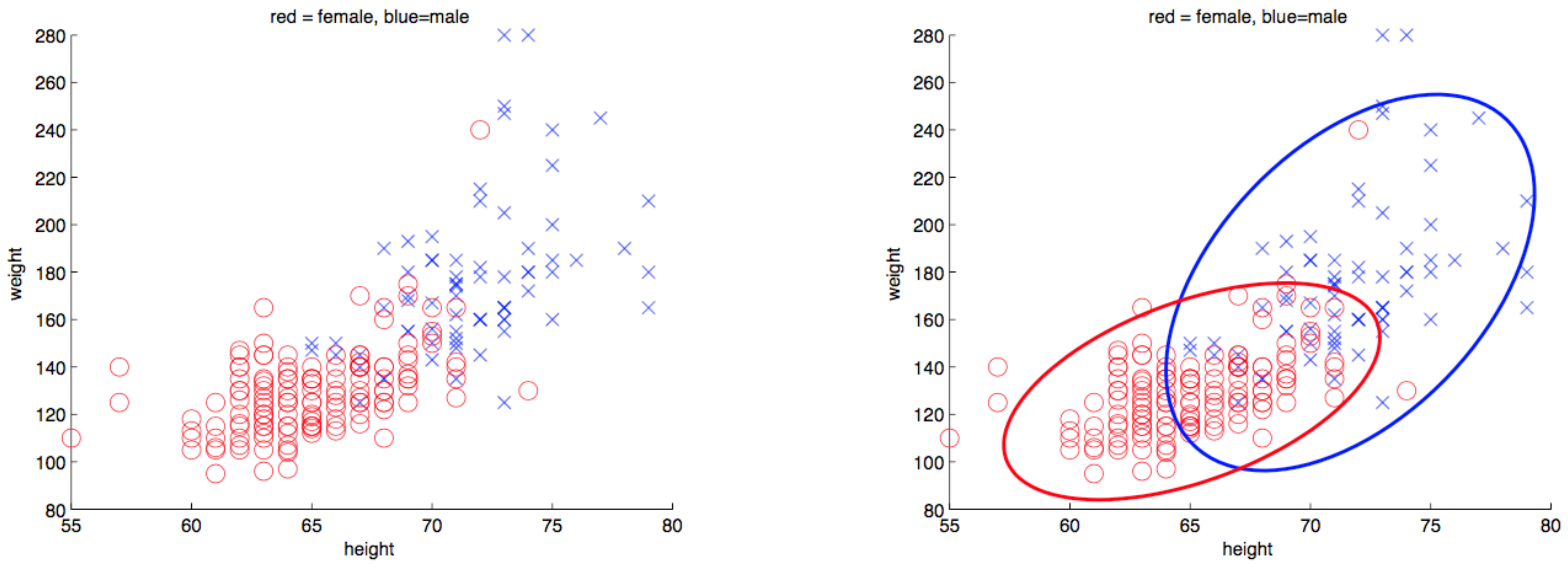


$$\Sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

[https://www.cs.ubc.ca/~murphyk/Teaching/CS340-Fall07/gaussClassif.pdf]

Gaussian Discriminant Analysis (GDA)

$$P(\mathbf{x} \mid \mathbf{y} = \mathbf{k}) = \mathcal{N}(\theta_{\mathbf{k}}, \Sigma_{\mathbf{k}})$$



With diagonal covariance (NB) the ellipses are **axis-aligned**

[Figure 4.2, Machine Learning. Kevin P. Murphy]

Linear Discriminant Analysis (LDA)

- GDA has many parameters (MxM per class)
  - More prone to overfit

Recall: X is M-dimensional
e.g., M=61,168 for 20-newsgroup

- An alternative is to model identical class covariance:

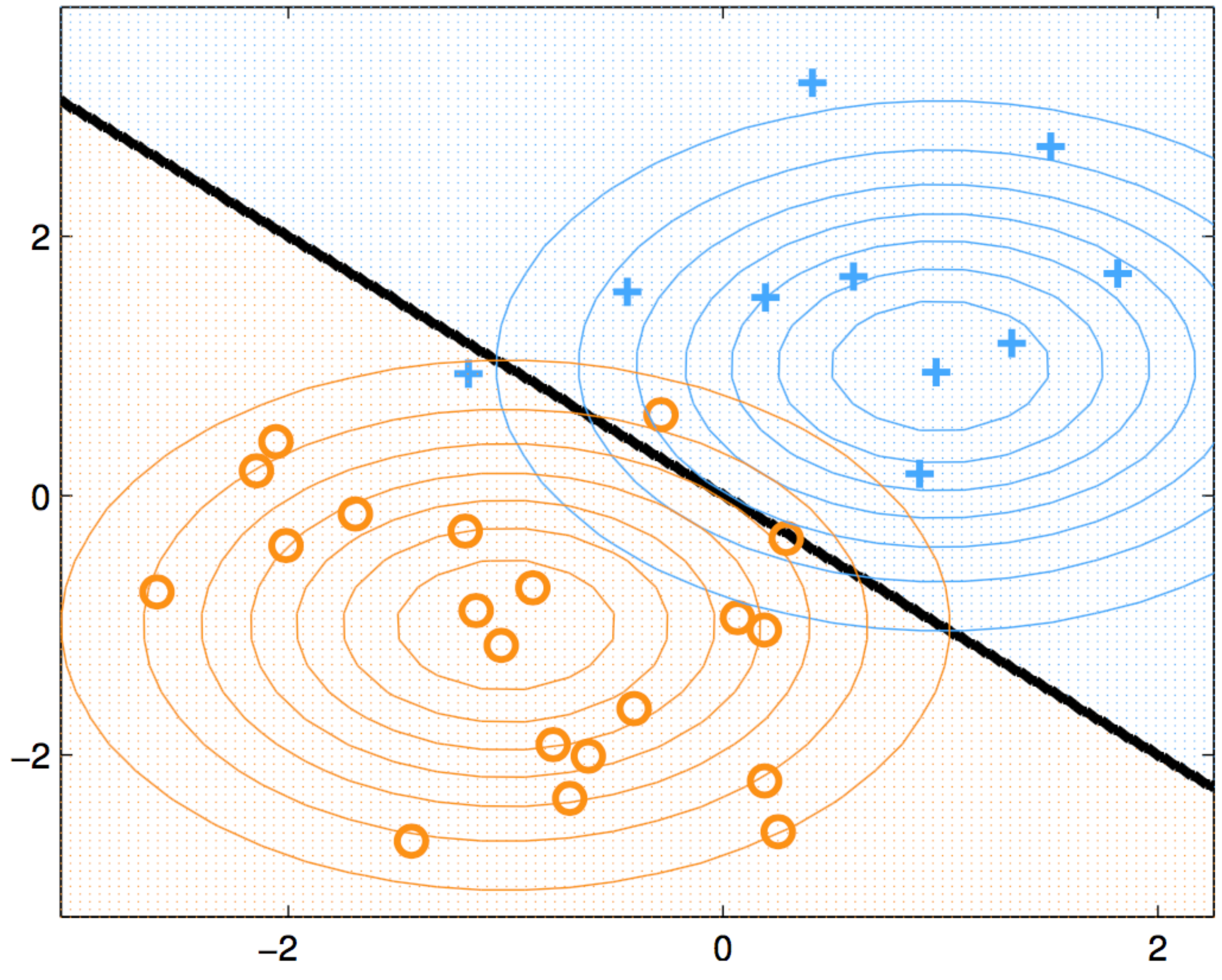
$$P(x \mid y = k) = \mathcal{N}(\theta_k, \Sigma_k)$$
$$\Sigma_k = \Sigma_{k'} \;\; \forall k, k'$$

- In the two class case the posterior over classes is similar to logistic regression:

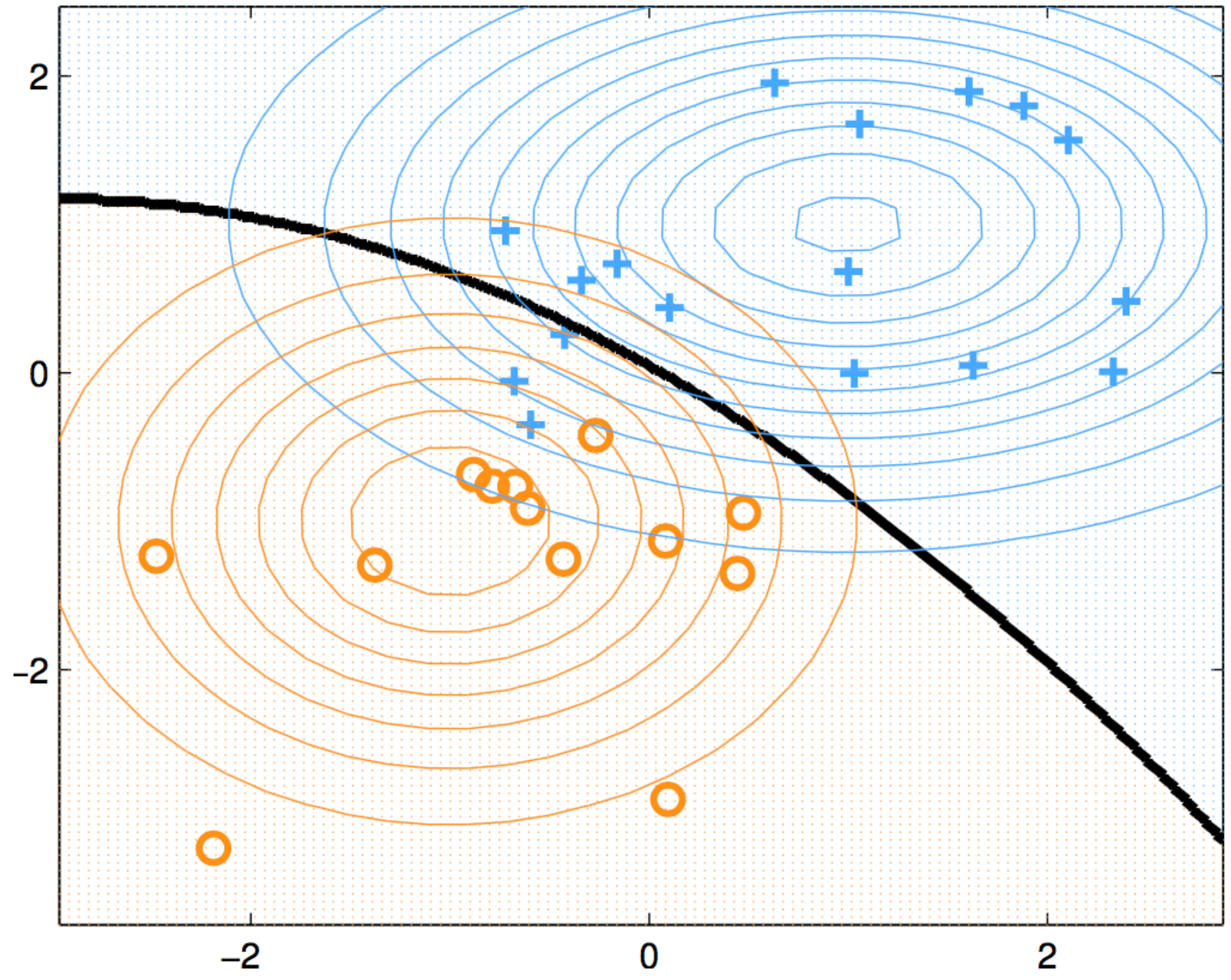
$$P(y = 1 \mid x, \theta) = \frac{1}{1 + \exp(f(x))}$$

Decision boundaries

Linear Discriminant Analys



Gaussian (Quadratic) Discriminant Analys

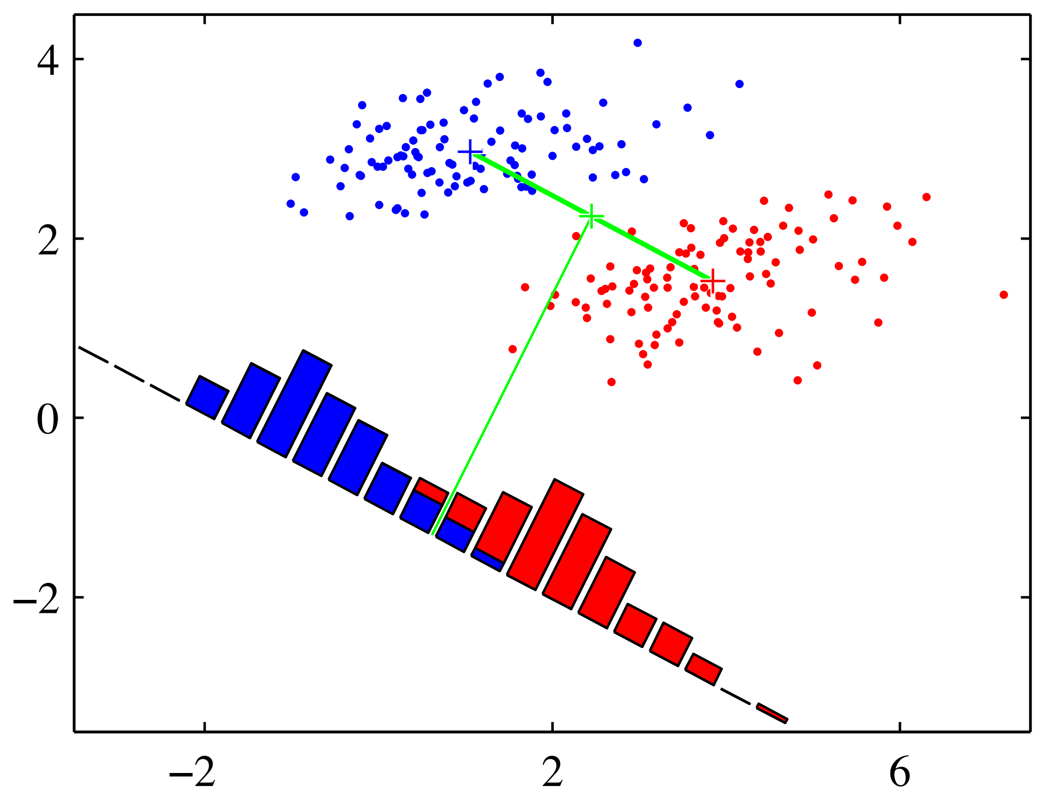


[Figure 4.3a & 4.5a, Machine Learning. Kevin P. Murphy]

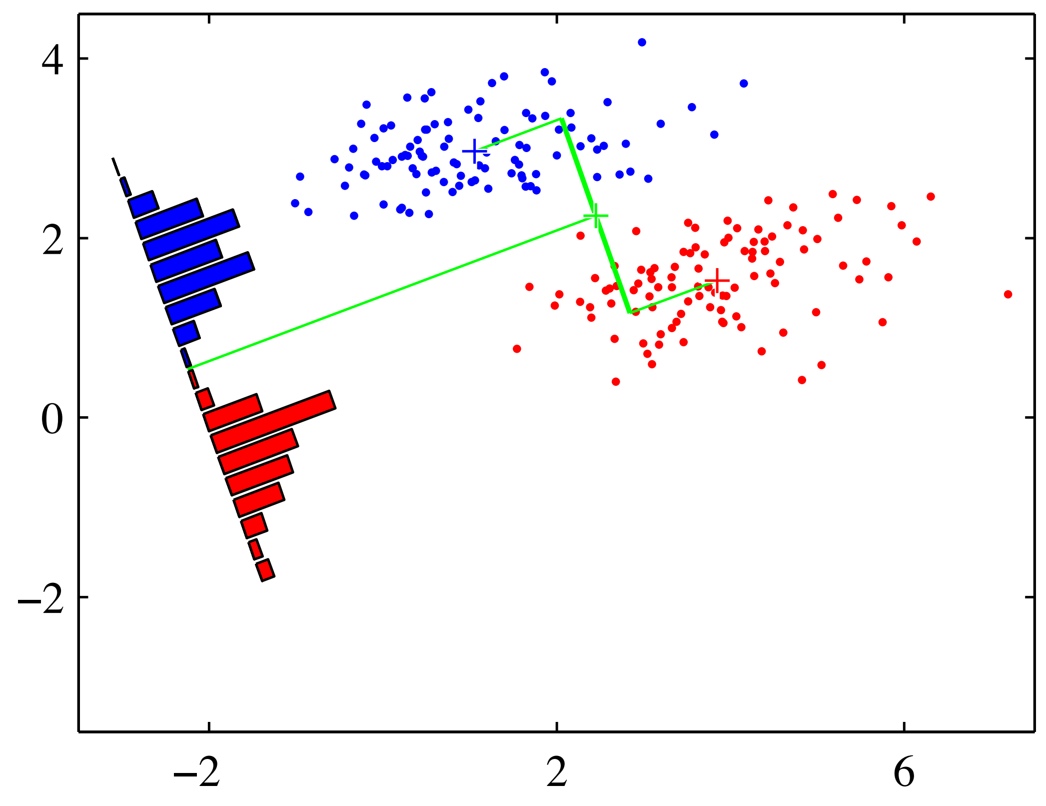
Fisher Discriminant Analysis (FLDA)

- **Reduce the dimensionality of the data (Wx) to obtain a linearly separable problem**

PCA



FLDA



Readings

- References
  - Sections 4.1—4.3, 4.5 of The Elements of Statistical Learning (available online)
  - Sections 3.5 & 4.2 of Machine Learning (K. Murphy)