CS480 – Introduction to Artificial Intelligence

Topic: Classification

Mustafa Bilgic

🔗 http://www.cs.iit.edu/~mbilgic

https://twitter.com/bilgicm

FEEDBACK

- Unsupervised learning
  - No feedback; the agent discovers patterns in the data
  - E.g., clustering, dimensionality reduction, outlier detection
- Supervised learning
  - Feedback: input-output pairs
  - E.g., classification, regression, ranking
- Reinforcement learning
  - Feedback: rewards

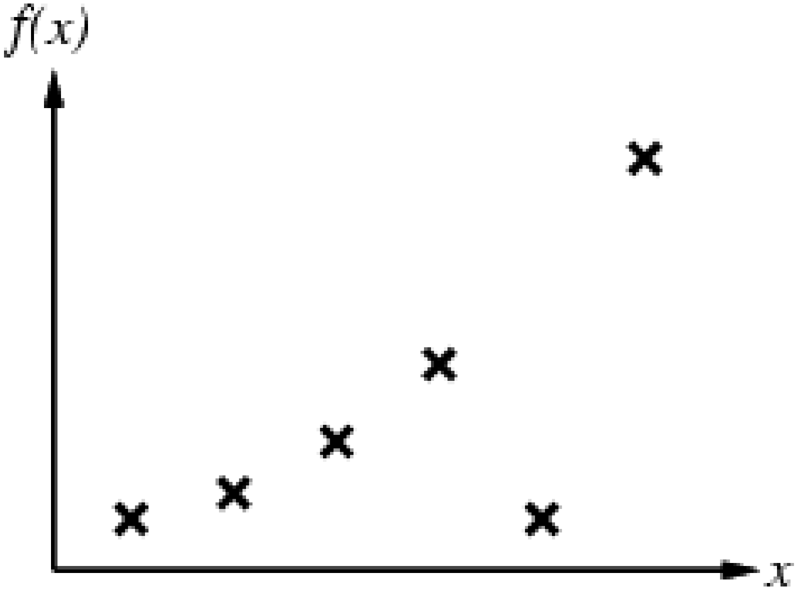
2. Supervised learning

- Given objects with their labels, <X,Y>
- Learn a function f that maps objects, X, to labels, Y
- We want f to perform well on unseen objects
- Several applications
  - Face recognition, speech recognition, medical diagnosis, fraud detection, credit scoring, home value prediction, temperature prediction, …
- If Y is
  - Discrete, the task is called classification
  - Continuous, the task is called regression

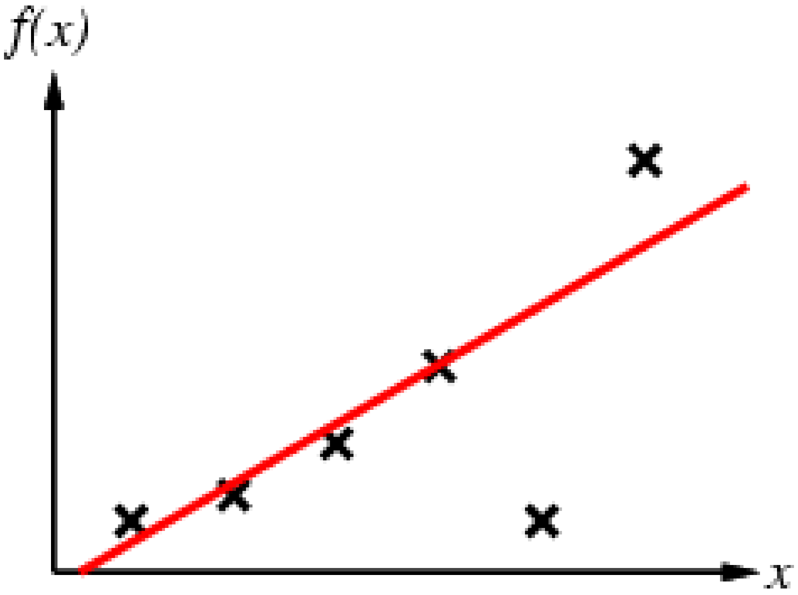
Function Fitting?

- Isn’t classification/regression simply “function fitting?”
- Yes and No
- The purpose is to generalize and perform well on unseen data
- We don’t want to underfit or overfit to the training data

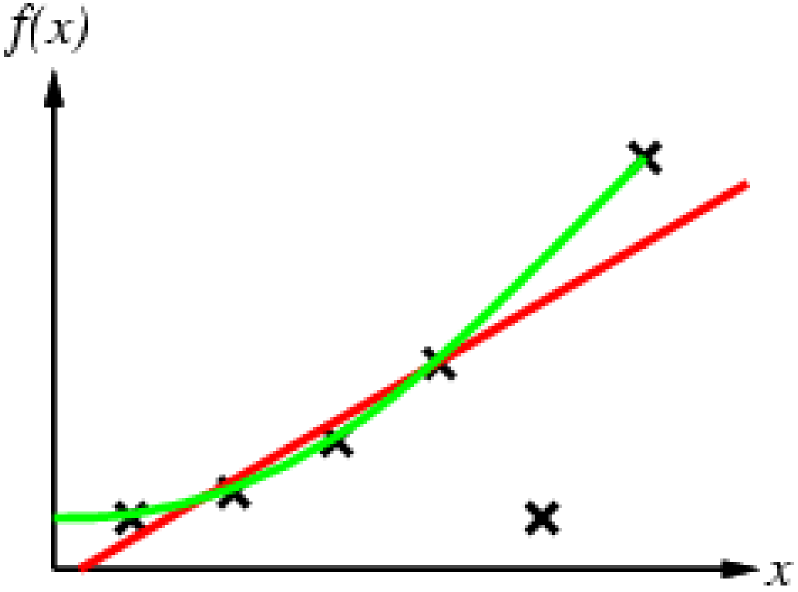
CURVE FITTING



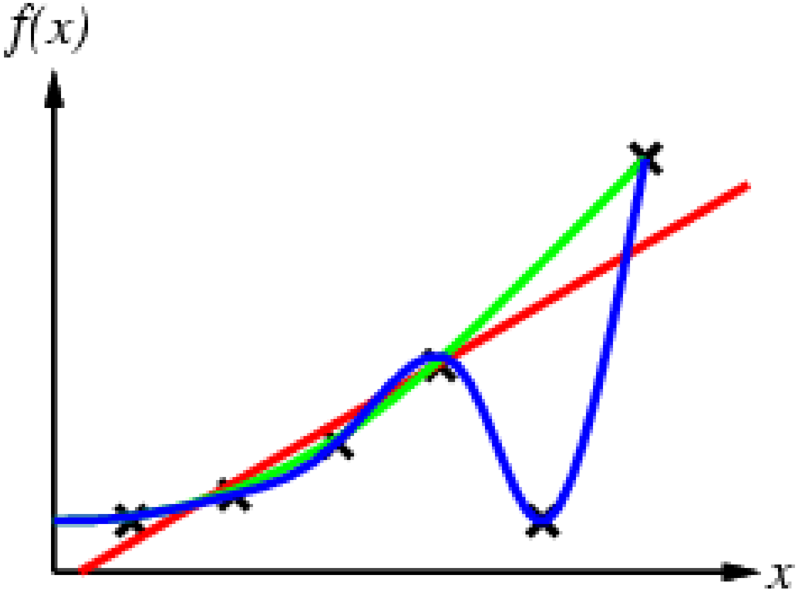
CURVE FITTING



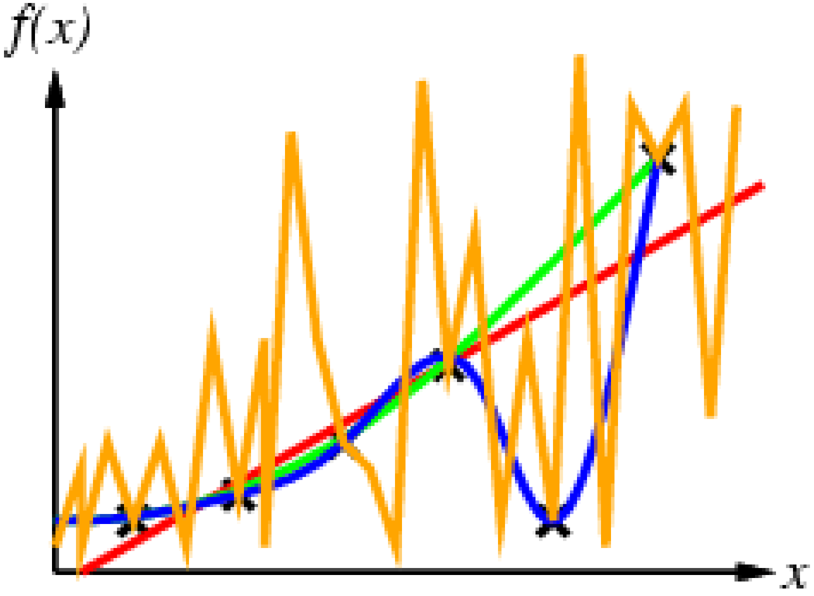
Curve Fitting



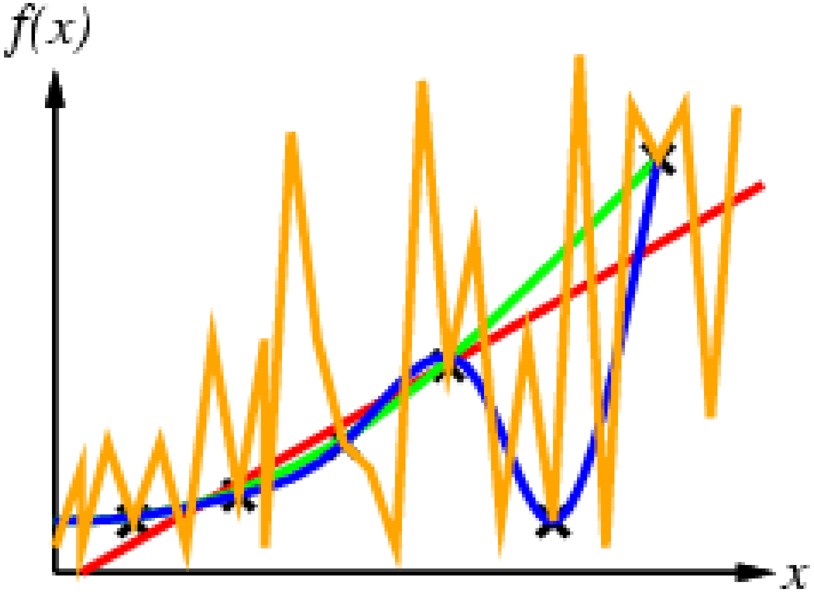
Curve Fitting



Curve Fitting



Curve Fitting



So, which function is the "right" one?

Classification Models

1. Decision trees
2. Nearest neighbors
3. Naïve Bayes
4. Logistic regression

   Note: it's called regression, but it is a classification model

5. Support vector machines
6. Neural networks

Naïve Bayes

TASK

- Classify emails as spam (s) / not-spam (~s) based on the words they contain
- You look at 100 random emails; 40 of them are spam, 60 of them are not-spam
- What is P(s) for a new email?

Features

- Assume you'll look into the emails' contents; you've decided that the word Nigeria[1] seems to correlate well with spam. You group the 100 emails as follows

Nigeria	Spam	Count
t	s	30
f	s	10
t	~s	10
f	~s	50

If the word Nigeria appears in the new email, then what is P(s | Nigeria=t)?

[1] Why "Nigeria?" https://www.google.com/search?q=nigeria+scam+emails

Nigeria=t

Nigeria	Spam	Count
t	s	30
f	s	10
t	~s	10
f	~s	50

If the word Nigeria appears in the new email, then what is P(s | Nigeria=t)?

$$P(s \mid N=t) = \frac{P(s, N=t)}{P(N=t)} = \frac{30/100}{(30+10)/100} = \frac{30}{40}$$

Add Admission into your vocabulary

Nigeria	Adm.	Spam	Count
t	t	s	10
t	f	s	20
f	t	s	3
f	f	s	7
t	t	~s	8
t	f	~s	2
f	t	~s	40
f	f	~s	10

What is P(s | N=t, A=f)? What about P(s | N=t, A=t)?

ADD ADMISSION INTO YOUR VOCABULARY

Nigeria	Adm.	Spam	Count
t	t	s	10
t	f	s	20
f	t	s	3
f	f	s	7
t	t	~s	8
t	f	~s	2
f	t	~s	40
f	f	~s	10

What is P(s | N=t, A=f)? What about P(s | N=t, A=t)?

$$P(s \mid N=t, A=f) = \frac{P(s, N=t, A=f)}{P(N=t, A=f)} = \frac{20/100}{(20+2)/100} = \frac{20}{22}$$

P(s | N=t) was 0.75. P(s | N=t, A=f) is 0.91

ADD ADMISSION INTO YOUR VOCABULARY

Nigeria	Adm.	Spam	Count
t	t	s	10
t	f	s	20
f	t	s	3
f	f	s	7
t	t	~s	8
t	f	~s	2
f	t	~s	40
f	f	~s	10

What is P(s | N=t, A=f)? What about P(s | N=t, A=t)?

$$P(s \mid N=t, A=t) = \frac{P(s, N=t, A=f)}{P(N=t, A=f)} = \frac{10/100}{(10+8)/100} = \frac{10}{18}$$

P(s | N=t) was 0.75. P(s | N=t, A=f) is 0.91. P(s | N=t, A=t) = 0.56.

Now assume we add 998 more words

W_1	W_2	…	W_{1000}	Spam	Count
t	t	…	t	s	
t	t	…	f	s	
…	…	…	…	…	
f	f	…	f	~s	

Q: How many entries are there in this table?

A: $2^{1001} \approx 2x10^{301}$

We have 100 emails. If all emails are distinct, 100 entries will be 1;
The rest will be 0.

Q: What is $P(s \mid W_1=t, W_2=f, \ldots, W_{1000}=t)$?

A: Either 1 or 0 if it is in D, otherwise, it is NaN

Q: How big of a training data do we need?

Naïve Bayes

- Given X_1, X_2, …, X_n, and class Y
- Assume $X_i \perp X_j \mid Y$

$$P(Y|X_1, X_2, \ldots, X_n) = \frac{P(X_1, X_2, \ldots, X_n|Y)P(Y)}{P(X_1, X_2, \ldots, X_n)} = \frac{P(Y)\prod_{i=1}^{n} P(X_i|C)}{P(X_1, X_2, \ldots, X_n)}$$

We need to estimate $P(X)$ and $P(X_i \mid C)$

What is the Bayesian network representation of Naïve Bayes?

Naïve Bayes

Nigeria	Adm.	Spam	Count
t	t	s	10
t	f	s	20
f	t	s	3
f	f	s	7
t	t	~s	8
t	f	~s	2
f	t	~s	40
f	f	~s	10

What is P(S)?

What is P(N|S)?

What is P(A|S)?

Naïve Bayes

Nigeria	Adm.	Spam	Count
t	t	s	10
t	f	s	20
f	t	s	3
f	f	s	7
t	t	~s	8
t	f	~s	2
f	t	~s	40
f	f	~s	10

What is P(S)?

Spam	P(S)
s	40/100
~s	60/100

What is P(N|S)?

Nigeria	Spam	P(N,S)	P(N|S)
t	s	30/100	30/40
f	s	10/100	10/40
t	~s	10/100	10/60
f	~s	50/100	50/60

What is P(A|S)?

Adm.	Spam	P(A,S)	P(A|S)
t	s	13/100	13/40
f	s	27/100	27/40
t	~s	48/100	48/60
f	~s	12/100	12/60

Inference in Naïve Bayes

- What is $P(s \mid N=t, A=f)$?

Zero probabilities

- We have n features, X_1 through X_n
- If $P(X_i|C)$ is zero for any feature and class combination, we would be in trouble
- Example
  - Assume that X_{592} is a weird feature that is rarely *true* in the world. Assume that X_{592} is always *false* in our training data, no matter what the class is
    - $P(X_{592} = f \mid C = t) = 1$; $P(X_{592} = t \mid C = t) = 0$
    - $P(X_{592} = f \mid C = f) = 1$; $P(X_{592} = t \mid C = f) = 0$
  - In one of the objects in our test data, X_{592} is *true*.
    - What is $P(C \mid X_1, X_2, \ldots, X_{592} = t, \ldots X_n)$?

Other Classifiers - Overview

Some Classifiers

- Naïve Bayes
- Logistic regression
- Decision trees
- Support vector machines
- Neural networks

Logistic Regression

- Learns $P(Y|\boldsymbol{X})$ directly, without going through $P(\boldsymbol{X}|Y)$ and $P(Y)$
- Assumes $P(Y|\boldsymbol{X})$ follows the logistic function

$$P(Y = false \mid X_1, X_2, \cdots, X_n) = \frac{1}{1 + e^{w_0 + \sum_{i=1}^{n} w_i X_i}}$$

$$P(Y = true \mid X_1, X_2, \cdots, X_n) = \frac{e^{w_0 + \sum_{i=1}^{n} w_i X_i}}{1 + e^{w_0 + \sum_{i=1}^{n} w_i X_i}}$$

- Learning: estimate the weights $w_0, w_1, \dots, w_n$

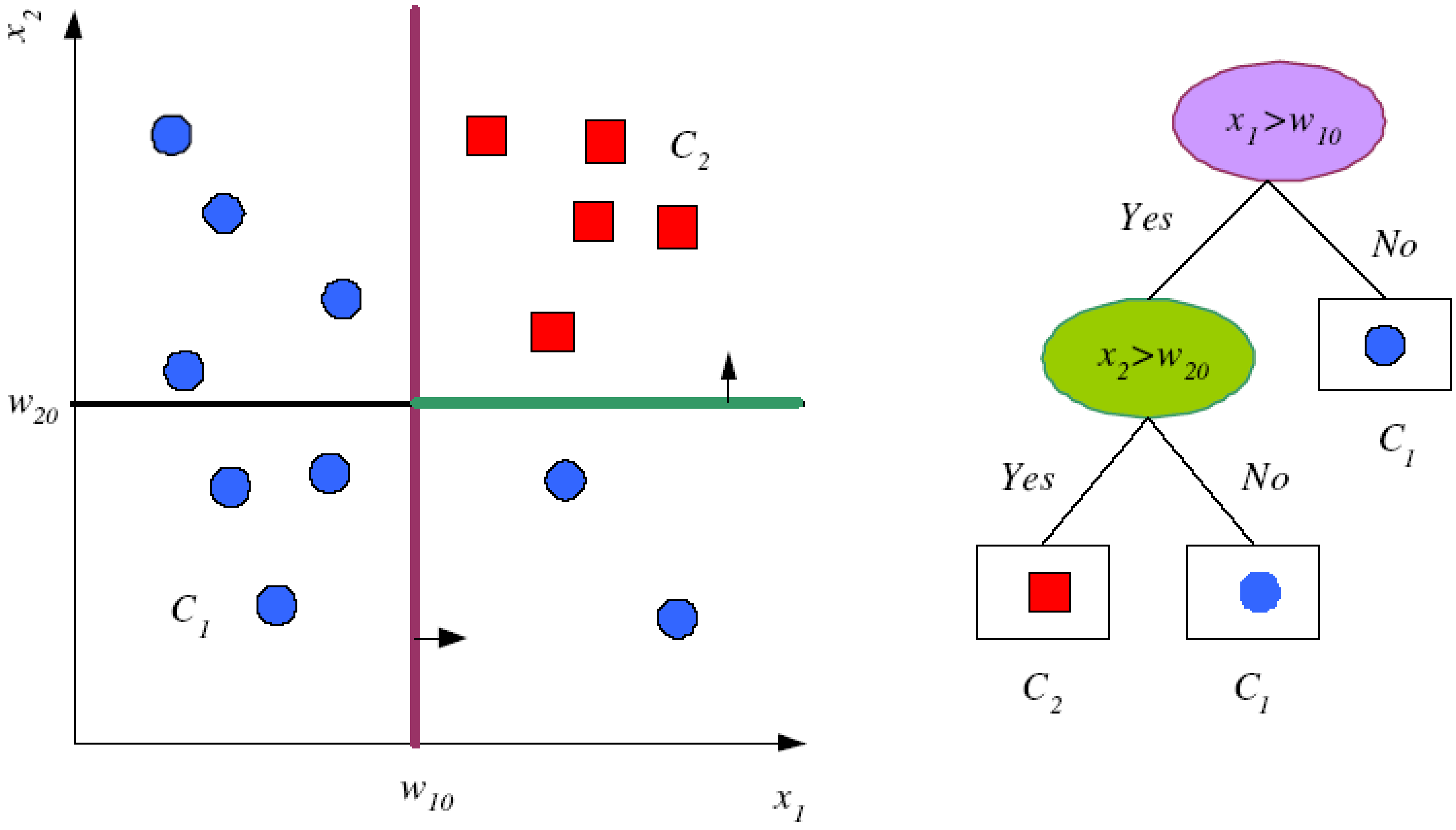
Learning – Parameter Estimation

- Maximize (conditional) log-likelihood

$$W \quad \leftarrow \quad \underset{W}{\operatorname{argmax}} \prod P(Y^{(d)} | \boldsymbol{X}^{(d)})$$

$$W \quad \leftarrow \quad \underset{W}{\operatorname{argmax}} \sum \ln P(Y^{(d)} | \boldsymbol{X}^{(d)})$$

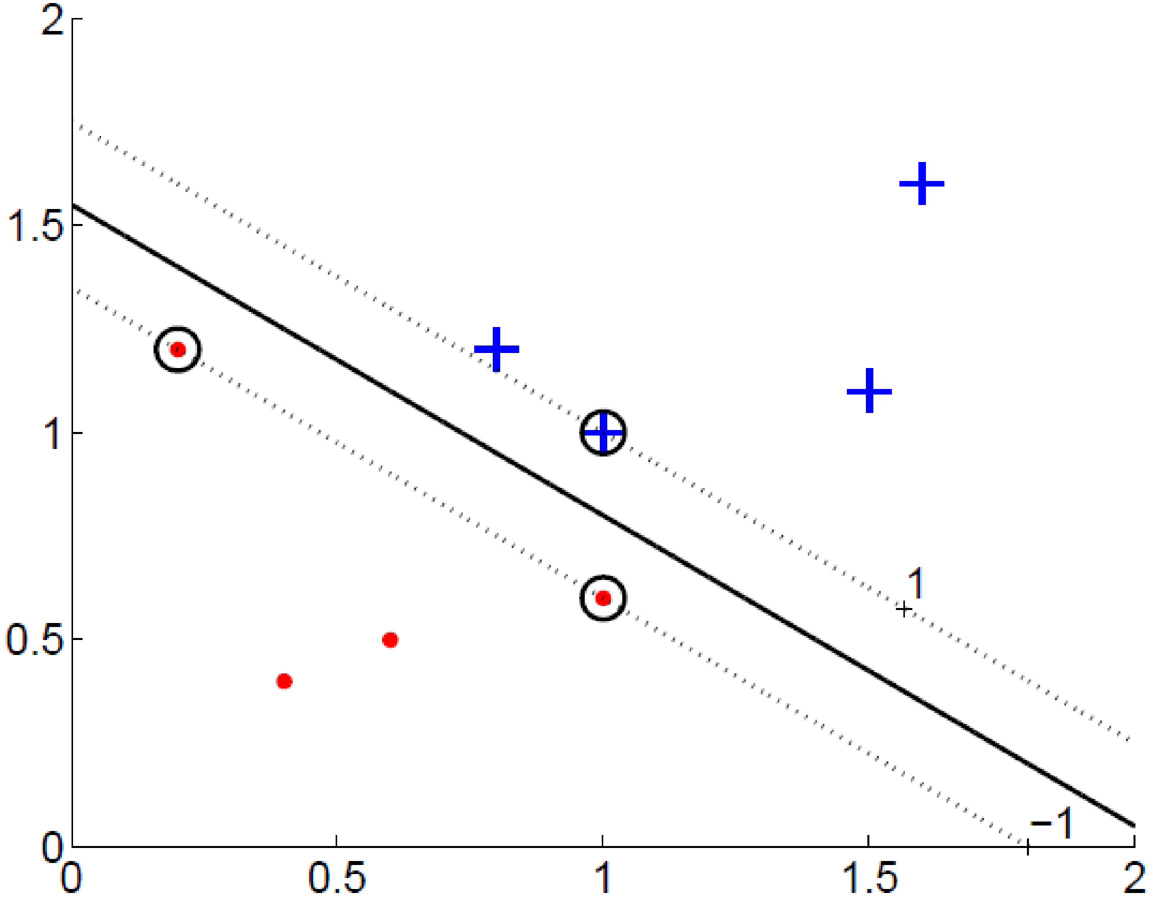
DECISION TREES



Learning: how do you learn a small tree that generalizes to unseen data?

Image credit: Ethem Alpaydin. Introduction to Machine Learning. 3rd Edition. http://www.cmpe.boun.edu.tr/~ethem/i2ml3e

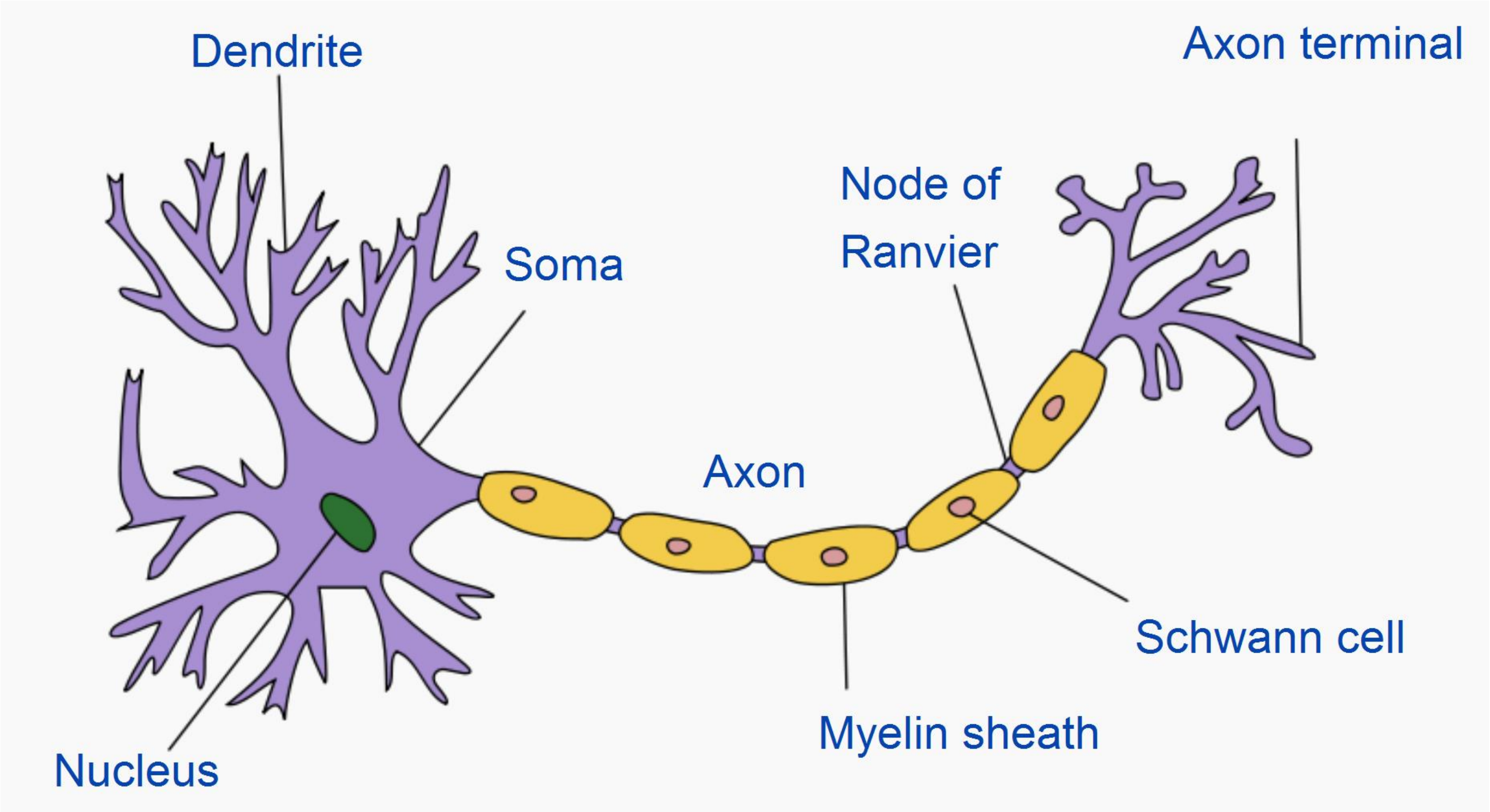
Support Vector Machines



$$\min \frac{1}{2} w^T w \text{ subject to } y^{(d)}\left(w^T x^{(d)} + b\right) \geq +1$$

Image credit: Ethem Alpaydin. Introduction to Machine Learning. 3rd Edition. http://www.cmpe.boun.edu.tr/~ethem/i2ml3e

NEURON



By Quasar Jarosz at English Wikipedia, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=7616130

What an Artificial Neuron Does

- Takes a weighted sum of its inputs
  - $w_0 + \sum_{i=1}^{k} w_i x_i$
  - Assume that there is always a constant input 1, that is, $x_0 = 1$. Then,
  - $\sum_{i=0}^{k} w_i x_i$
- Passes this sum through its activation function
  - $f\left(\sum_{i=0}^{k} w_i x_i\right)$

Multilayer Neural Networks

- An input layer
- One or more hidden layers
- An output layer

- Learning: estimate the weights

Scikit-learn Code Examples

- https://scikit-learn.org/stable/
- Naïve Bayes
  - https://scikit-learn.org/stable/modules/naive_bayes.html
- Logistic regression
  - https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
- Decision Trees
  - https://scikit-learn.org/stable/modules/tree.html
- Support vector machines
  - https://scikit-learn.org/stable/modules/svm.html
- Neural networks
  - https://scikit-learn.org/stable/modules/neural_networks_supervised.html