Strengths and weaknesses of different machine learning models

APSC 8280: Machine learning applied to plant science

03/26/2020

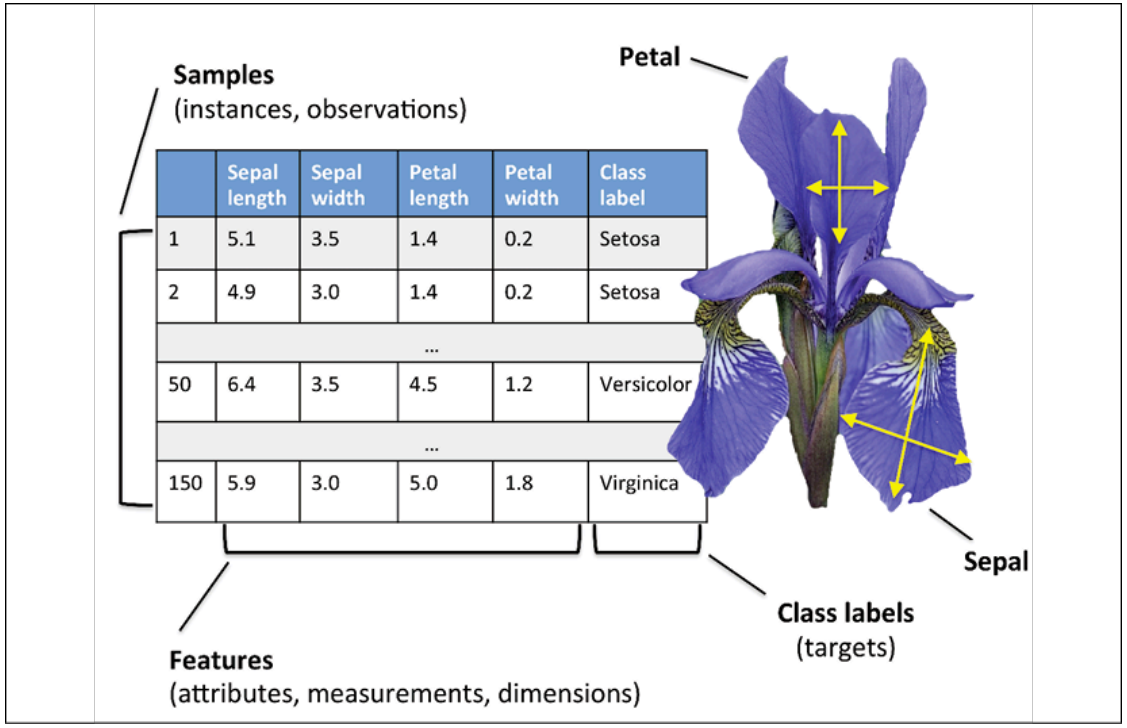
Outline

- Review lab 1
- Demo 1: End-to-end machine learning
- Quiz on machine learning fundamentals
- Algorithms
  - Naïve Bayes
  - Regression techniques
  - Support vector machines
  - Nearest neighbors
  - Decision trees
  - Ensemble methods

Demo: End-to-End Machine Learning

Datasets

Classification: flower identification



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

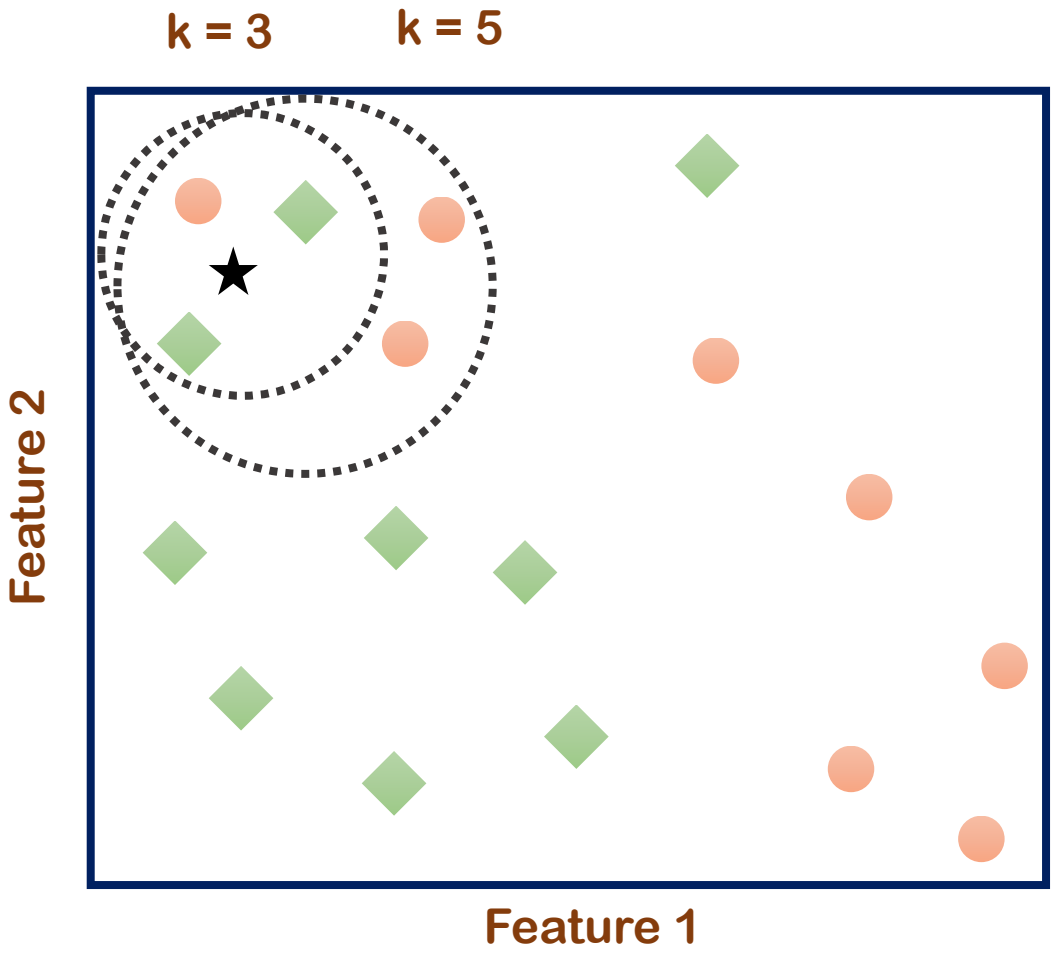
Regression: House prices prediction

Raschka, Sebastian, and Vahid Mirjalili. Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2. Packt Publishing Ltd, 2019.
https://towardsdatascience.com/machine-learning-project-predicting-boston-house-prices-with-regression-b4e47493633d

Nearest neighbors

Plant disease prediction

Healthy Diseased



- Classifies unlabeled examples by assigning them the class of similar labeled examples.
- Best for data with numerous complex relationships among features and target variables
- Rule of thumb: start with k equal to the square root of the number of training examples
- Large number of training examples makes the choice of k less important
- Heavily affected by the scale of the data

Nearest neighbors

Strengths

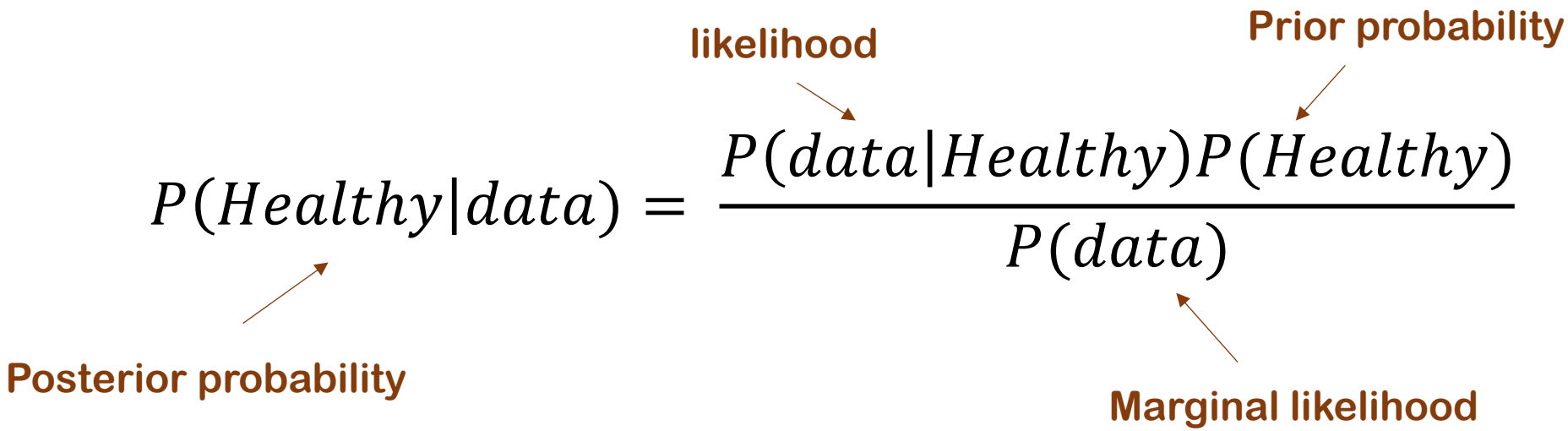
- ✓ Simple and effective
- ✓ Easy to train
- ✓ Makes no assumptions

Weaknesses

- X Does not produce a model (instance-based)
- X Hard to test
- X Struggles nominal features and missing data

Naïve Bayes

- Naïve Bayesian classifiers compute probability of each class using feature values from training data
- Best applied to problems in which information from many features should be considered simultaneously in order to estimate the probability of an outcome.
- Assumes all features are equally important and independent



Naïve Bayes

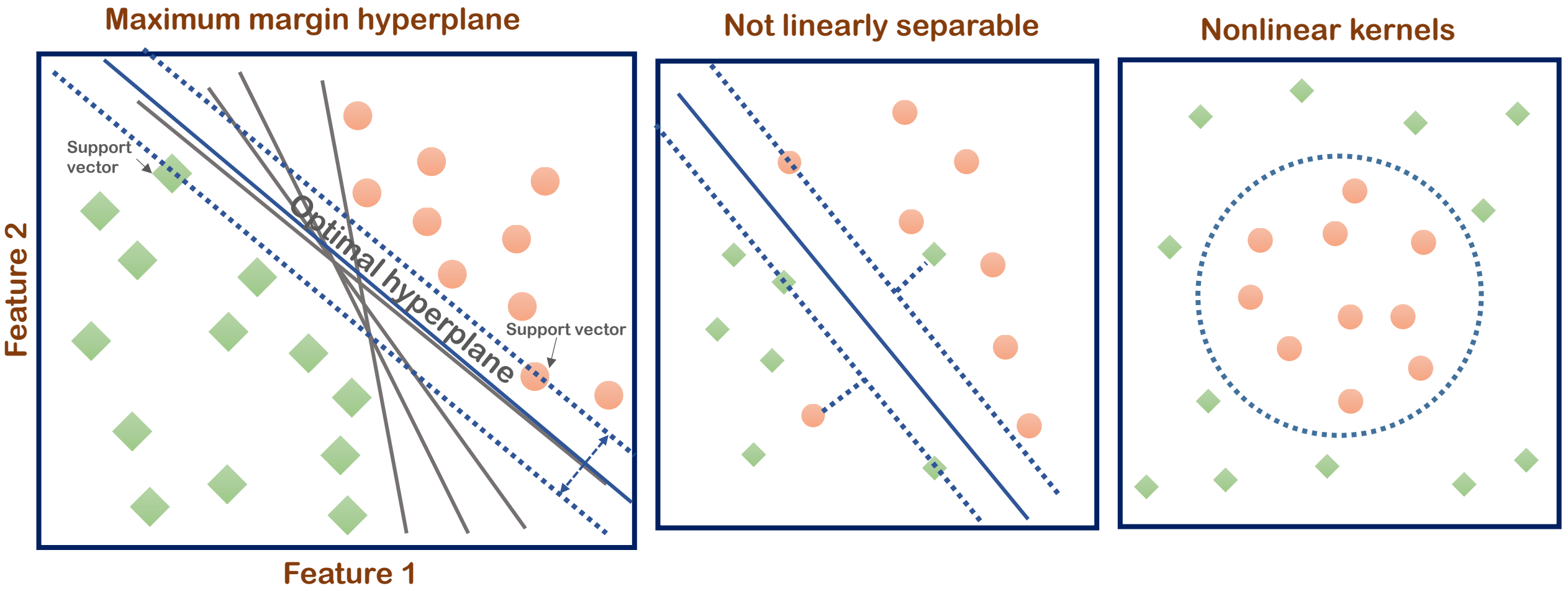
Strengths

- ✓ Simple, fast and effective
- ✓ Handles noisy and missing data well
- ✓ Requires relatively few examples for training
- ✓ Easy to obtain the estimated probability for a prediction

Weaknesses

- X Relies on the assumption of equally important independent features
- X Not ideal for datasets with many numeric features
- X Estimated probabilities are less reliable than the predicted classes

Support vector machines



- SVMs use multidimensional surfaces to define relationships between features and outcomes
- Support vectors are data points from each class that are closet to the MMH.
- They combine aspects of instance-based methods and linear regression
- Uses the concept of maximum margin hyperplane: greatest separation between the classes

Support vector machines

Strengths

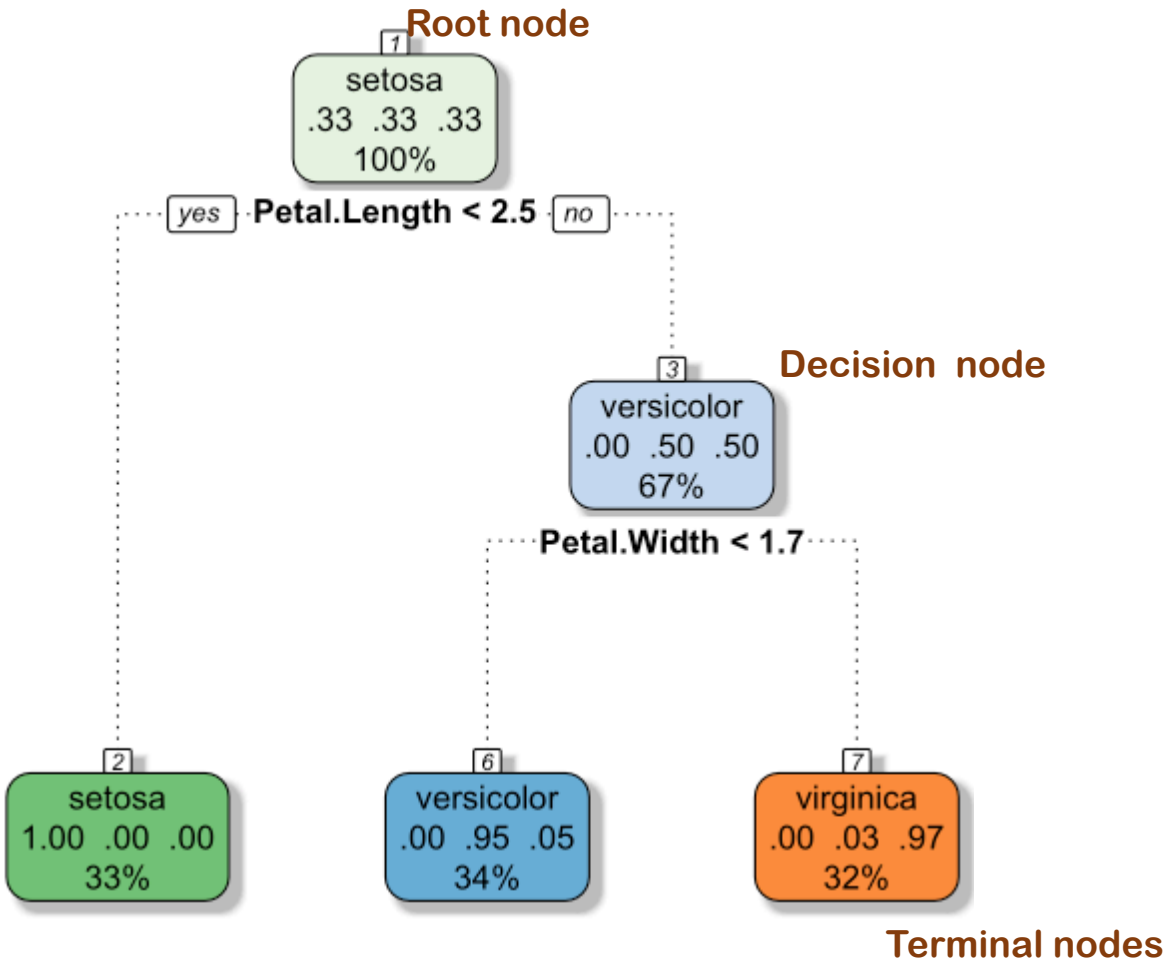
- ✓ Can be used for classification or regression
- ✓ Not overly influenced by noisy data and not very prone to overfiting
- ✓ Popularity due to high profile wins in data mining competitions

Weaknesses

- X So many options (kernels and model parameters)
- X Slow to train
- X Complex black box model

Decision trees

- DTs use a tree structure to model the relationships among the features and target variable
- Based on recursive partitioning (divide and conquer)
- Stopping criterion:
  - Homogenous
  - No more features to distinguish examples
  - The tree has grown to a predefined size limit
- Human readable format
- Not an ideal fit for large number of nominal features with many levels or many numeric features. Why?



Decision trees

Strengths

- ✓ Wide applicability
- ✓ Can handle numeric or nominal features, missing data
- ✓ Excludes unimportant features
- ✓ Works well with both small and large datasets
- ✓ Readily interpretable
- ✓ More efficient than complex models

Weaknesses

- X Typically biased towards splits on features with large number of levels
- X Very easy to overfit or underfit
- X Can have trouble modeling complex relationships
- X Small changes in training data can result in large changes in logic
- X Large trees are difficult to interpret

Regression techniques

Regression involves specifying a relationship between a dependent variable and one or more independent variables

$$\hat{y} = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

Parameters

Hyperparameters

$$RSS = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2$$

$$+ \lambda \left[\frac{(1-\alpha)}{2} \sum_{i=1}^{p} \beta_i^2 + \alpha \sum_{i=1}^{p} |\beta_i| \right]$$ **Elastic net**

Classification: logistic regression

$$z = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

$$p = \frac{1}{1 + e^{-z}}$$

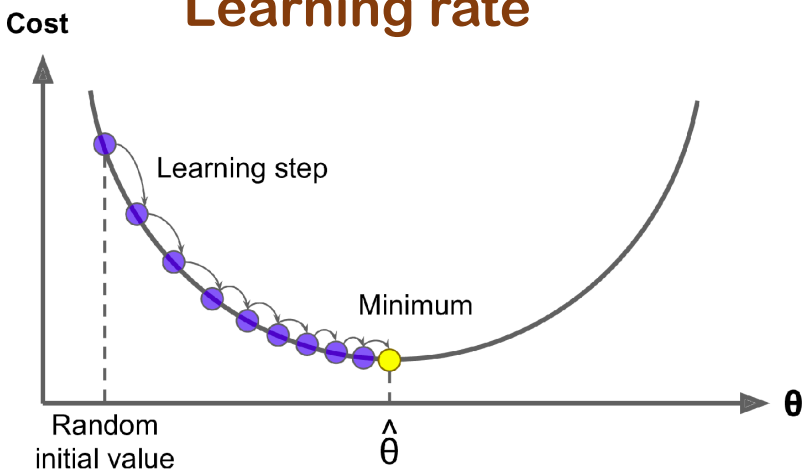
$$\hat{y} = \begin{cases} 0 \; if \; p < 0.5 \\ 1 \; if \; p \geq 0.5 \end{cases}$$

Multi-class classification

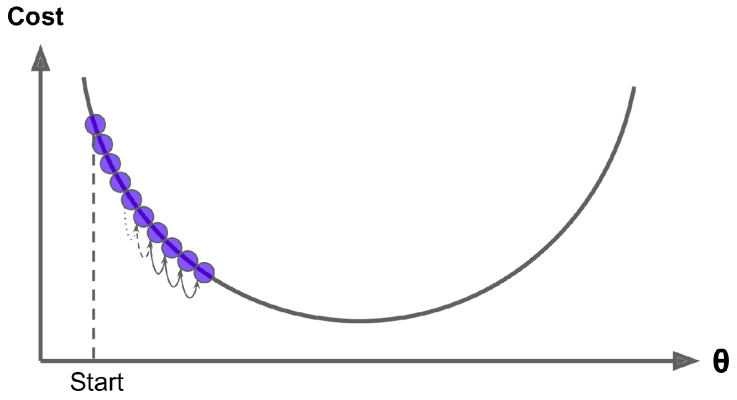
$$p_k = \frac{\exp(z_k)}{\sum_{k=1}^{K} \exp(z_k)}$$

Gradient descent and learning rates

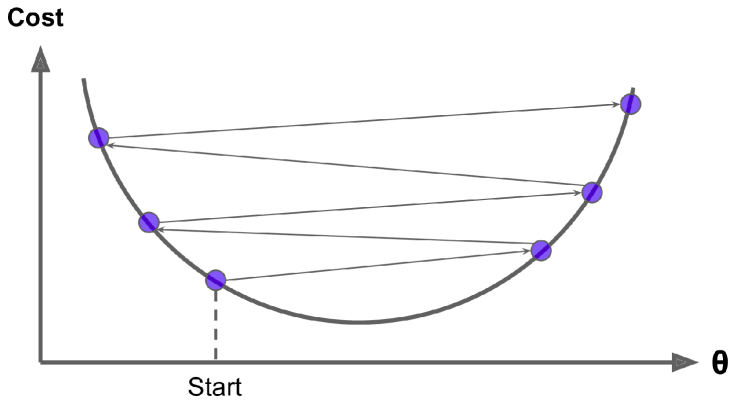
Learning rate



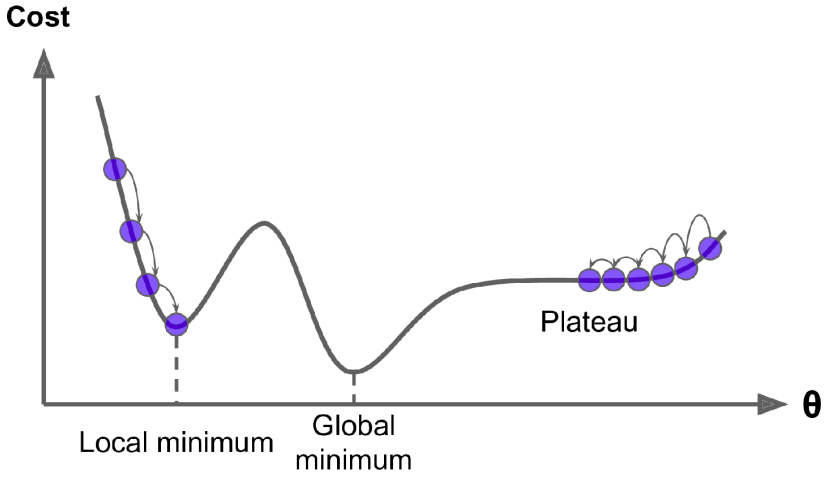
Learning rate too small



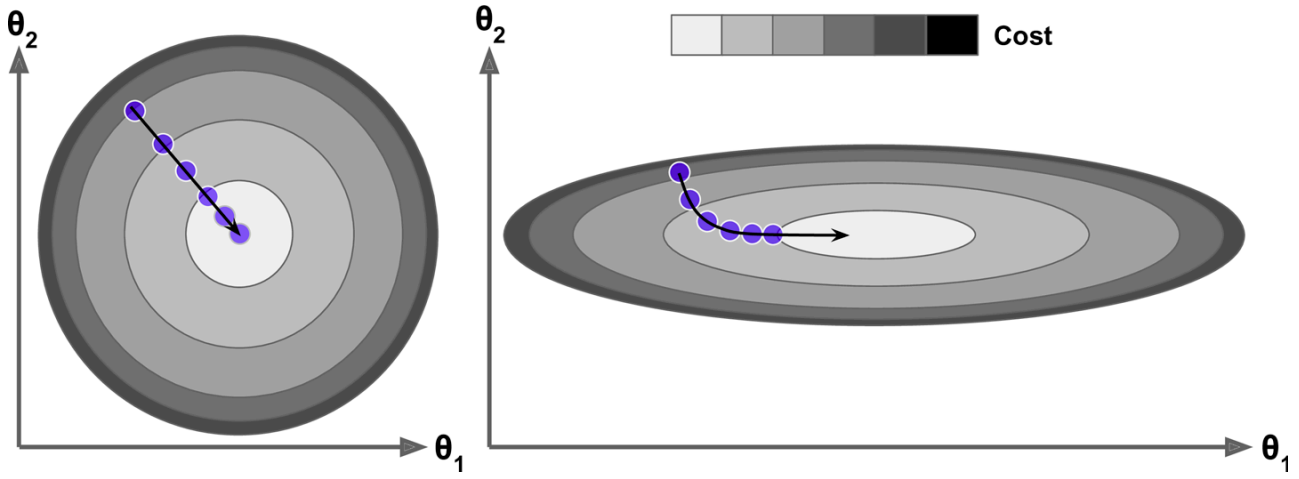
Learning rate too large



Global and local minimums



Impact of feature scaling



Regression techniques

Strengths

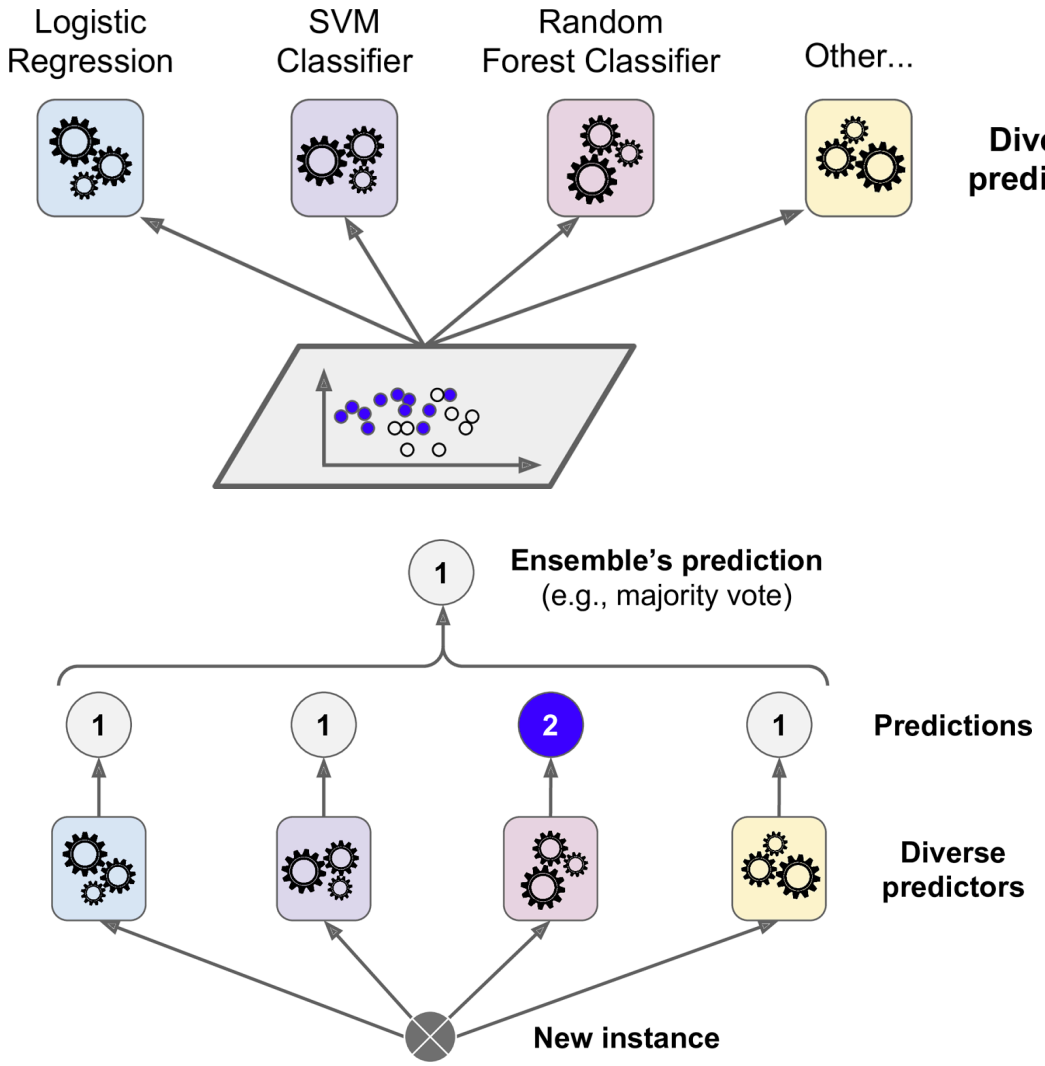
- ✓ By far the most common approach for modeling numeric data
- ✓ Can be adapted to model any task
- ✓ Provides estimates of the size and strength of relationships among features and the outcome

Weaknesses

- X Makes strong assumptions about the data
- X Model's form must be specified in advance
- X Cannot handle missing data
- X Works with only numeric features
- X Requires good knowledge of statistics to use/understand the models properly

Ensemble methods

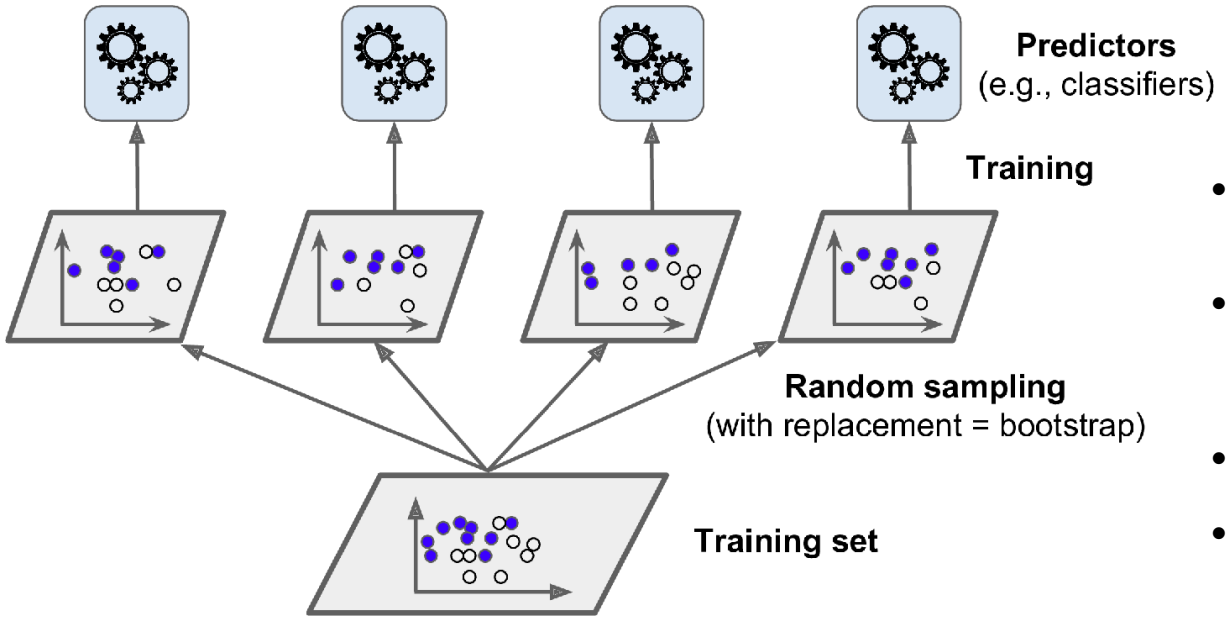
Voting classifiers

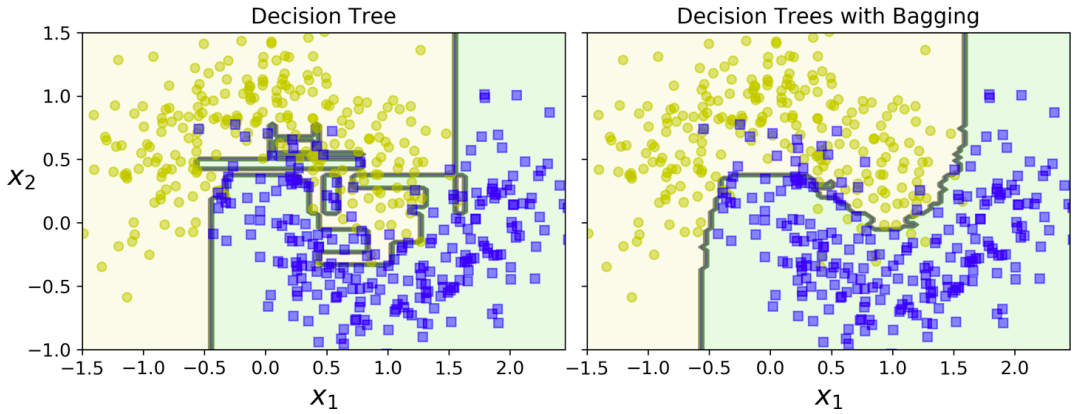


- Voting often achieves higher accuracy than the best classifier in the ensemble.
- Learners need to be diverse
- Hard voting uses class labels, soft voting using class probabilities
- Different kinds of voting schemes can be devised

Ensemble methods

Booststrap aggregating (Bagging)

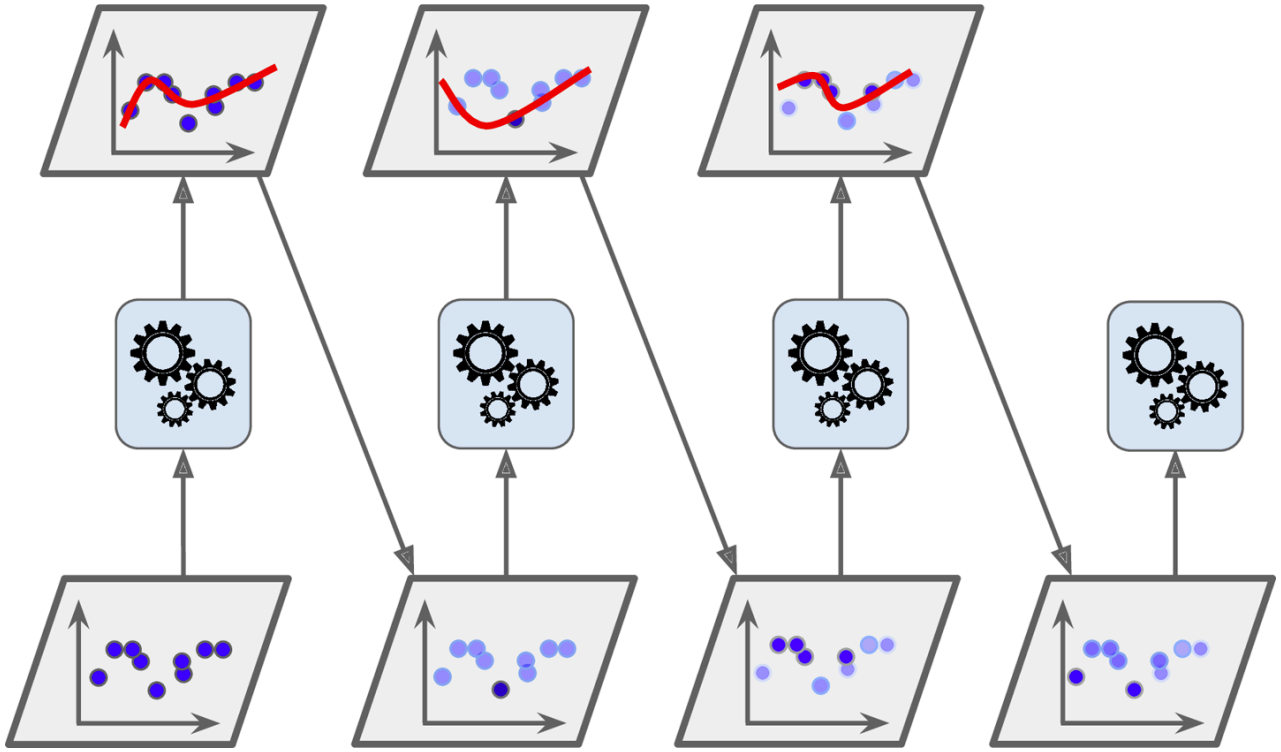




- Uses the same training model but different random subsets of the data.
- Sampling with replacement: bagging, without replacement: pasting
- Out of bag instances can be used as a validation set
- Random forest is a bagging of decision trees with random subset of features used at each node of the tree
- Extremely randomized trees (extra-trees) use random thresholds at each node

Ensemble methods

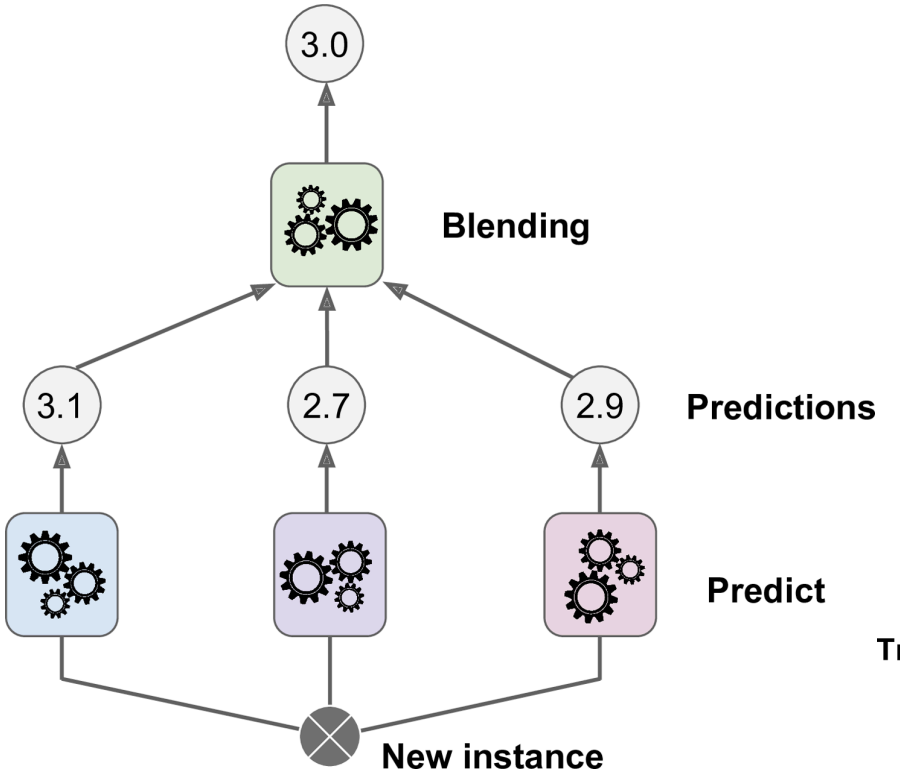
Boosting

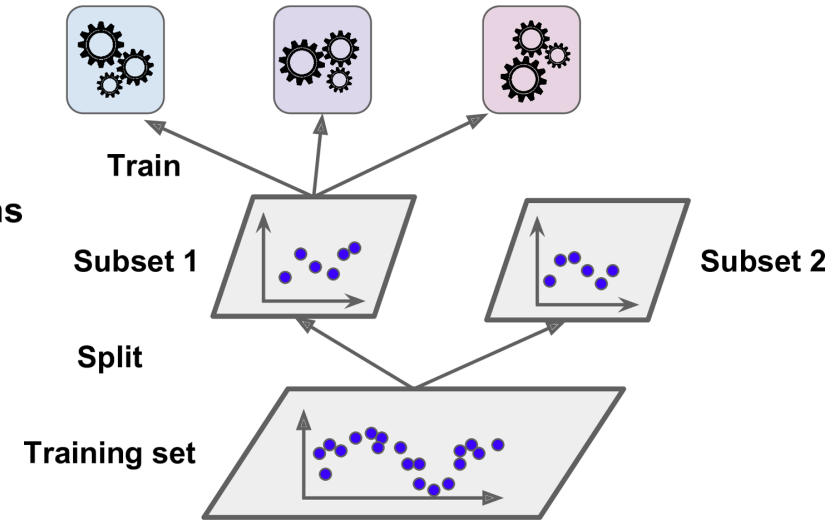


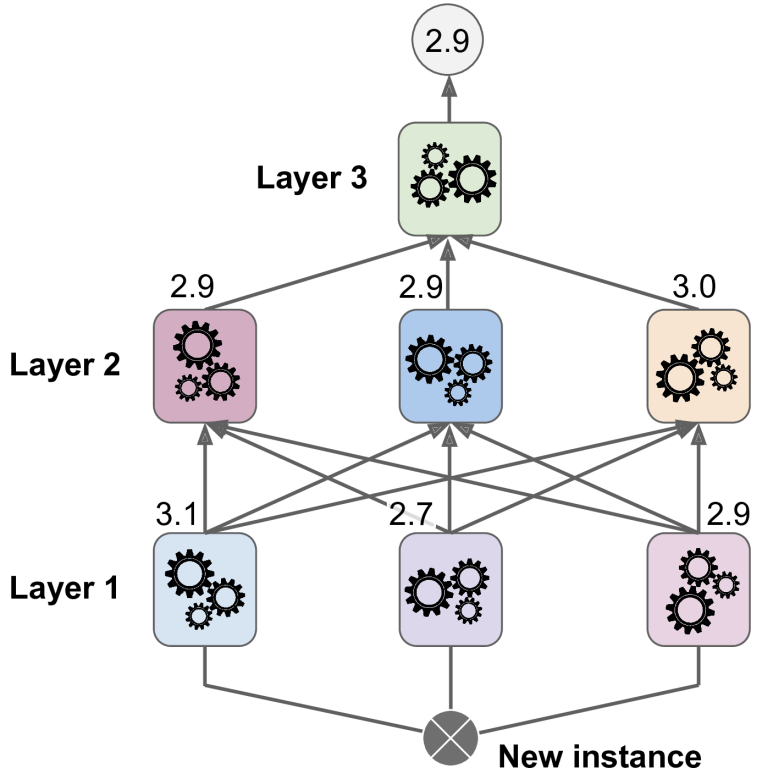
- Training learners sequentially
- The new predictor tries to correct the errors its predecessor
- Pay more attention to the training instances that the predecessor underfitted (AdaBoost)
- In gradient boosting, the new predictor is fitted to the residual errors made by the previous predictor

Ensemble methods

Stacking







Ensemble methods

Strengths

- ✓ Improved performance
- ✓ Models typically have lower variances and are more robust
- ✓ Can capture simple and complex relationships in the data

Weaknesses

- X Can be difficult to train
- X Much larger parameter space
- X Reduced model interpretability

Discussion questions

1. How would you define Machine Learning?
2. What are the two most common supervised tasks?
3. What type of Machine Learning algorithm would you use to allow a robot to walk in various unknown terrains?
4. What type of algorithm would you use to segment your customers into multiple groups?
5. What type of learning algorithm relies on a similarity measure to make predictions?
6. What is the difference between a model parameter and a learning algorithm's hyperparameter?
7. What do model-based learning algorithms search for? What is the most common strategy they use to succeed? How do they make predictions?
8. If your model performs great on the training data but generalizes poorly to new instances, what is happening? Can you provide three possible solutions?
9. What is a test set, and why would you want to use it?
10. What is the purpose of a validation set?
11. What can go wrong if you tune hyperparameters using the test set?
12. What is an online learning system?