Machine Learning & Deep Learning

Reda Khoufache & Bilal FAYE

Master 2 MIAGE & BUT3 Informatique

Université Paris Descartes & UIT Villetaneuse

2023/2024

Outline

Introduction

Motivation

Course organization

Outline

Introduction into mathematical modeling

Consider predicting a person's height from his age. Mathematically, this task can be written as follows:

$$height = f(age)$$

where f is a logarithmic function with parameters:

$$f(age) = 2.2 \log(80 \times age).$$

In model $f(age) = 2.2 \log(80 \times age)$:

- Both input and output (age and height) are single numbers
- In the case where the input and output are vectors, the model would have been expressed with matrices
- **Linear Algebra** is the branch of mathematics that study linear transformations between vectors in a linear fashion
- In Machine Learning or Deep Learning, models are built over 4 algebraic objects:
  1. Scalars
  2. Vectors
  3. Matrices
  4. Tensors

Scalars

Scalars are single numbers defined by the set to which they belong. For example, number of persons in the classroom is $p \in \mathbb{N}$ while the temperature is $t \in \mathbb{R}$. In Python, scalars are defined as:

```
n = 52
t = 12.5
print("we have " + str(n) + " students")
print("and the temperature is " + str(t) + " degrees.")
```

Vectors

A vector is an ordered array of numbers. Let $\mathbf{x}$ be a vector with d elements

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}.$$

If $x_1, \ldots, x_d \in \mathbb{R}$ then we write $\mathbf{x} \in \mathbb{R}^d$; we say that $\mathbf{x}$ is a d-dimensional vector or a vector of dimension d.

Vectors

Let $\mathbf{x}$ and $\mathbf{y}$ be two vectors defined in $\mathbb{R}^2$ such that

$$\mathbf{x} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}, \ \mathbf{y} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}.$$

$\mathbf{x}$ and $\mathbf{y}$ are geometrically represented as follows

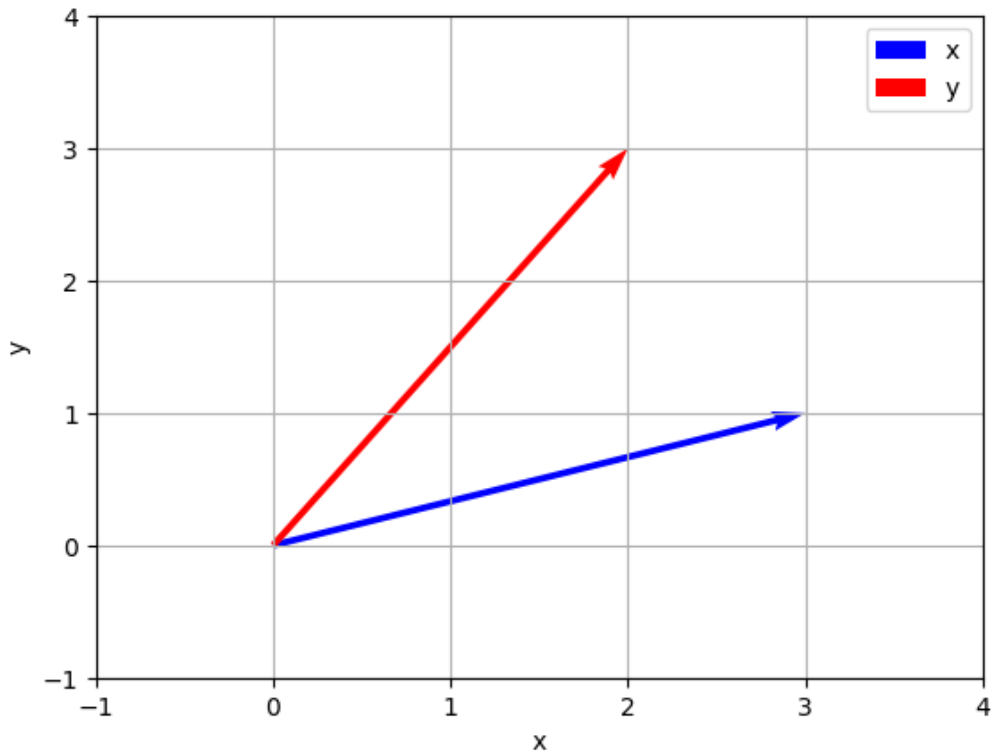

Figure: Geometrical representation of $\mathbf{x}$ and $\mathbf{y}$

Vectors

Vectors are essentially characterized by:

1. Direction
2. Length obtained with the Pythagorean theorem

$$\|\mathbf{x}\| = \sqrt{\sum_{i=1}^{d} x_i^2}.$$

- Vectors are objects characterized by direction and length
- Points have no direction, and their length is equal to zero

We tend to represent vectors as points for readability but remember these points are related to the origin and possess direction and length.

Vectors: Transposition

Definition

Let $\mathbf{x} \in \mathbb{R}^d$ be vector, by default, $\mathbf{x}$ is defined as column vector

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}.$$

$\mathbf{x}$ transpose is written $\mathbf{x}^T$, and is defined by

$$\mathbf{x}^T = [x_1, \ldots, x_d].$$

Then, the column vector $\mathbf{x}$ can be written as a row vector through the transpose operator:

$$\mathbf{x} = [x_1, \ldots, x_d]^T$$

Column and row versions of the same vector differ and relate through the transpose operator.

Vectors: Scalar multiplication

Definition

Let $\alpha \in \mathbb{R}$ be a scalar and $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{x}$ can be scaled by α as follows

$$\alpha\mathbf{x} = \begin{bmatrix} \alpha x_1 \\ \vdots \\ \alpha x_d \end{bmatrix} \in \mathbb{R}^d.$$

Example: Let $\alpha = 2$, $\beta = -\frac{1}{2}$ be two scalars, and $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^2$ two 2-dimensional vectors given by

$$\mathbf{x} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}.$$

Hence, the scaled vectors by α and β respectively are

$$\alpha\mathbf{x} = \begin{bmatrix} 6 \\ 2 \end{bmatrix}, \quad \beta\mathbf{y} = \begin{bmatrix} -1 \\ -\frac{3}{2} \end{bmatrix}.$$

Vectors: Unit vector

Definition

Let $\mathbf{x} \in \mathbb{R}^d$. The unit-vector of $\mathbf{x}$ is $\mathbf{u}$ defined as

$$\mathbf{u} = \frac{1}{\|\mathbf{x}\|}\mathbf{x}$$

$\mathbf{u}$ has same direction as $\mathbf{x}$ with $\|\mathbf{u}\| = 1$.

Vectors: Addition

Definition

Let $\mathbf{x}$ and $\mathbf{y} \in \mathbb{R}^d$. Addition between $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$\mathbf{x} + \mathbf{y} = \begin{bmatrix} x_1 + y_1 \\ \vdots \\ x_d + y_d \end{bmatrix} \in \mathbb{R}^d$$

The addition of two vectors and multiplication by scalars must always produce vectors, i.e., if $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ two vectors and $\lambda \in \mathbb{R}$ a scalar, then we have

$$\mathbf{x} + \lambda\mathbf{y} \in \mathbb{R}^d$$

Vectors: Dot product

Definition

Dot product between $\mathbf{x}$ and $\mathbf{y} \in \mathbb{R}^d$ is scalar defined as

$$\mathbf{x} \cdot \mathbf{y} = \|\mathbf{x}\| \times \|\mathbf{y}\| \times \cos(\theta),$$

where θ is the angle between $\mathbf{x}$ and $\mathbf{y}$;
$\cos(\theta) \in [-1, 1]$ measures how much $\mathbf{x}$ and $\mathbf{y}$ are oriented in terms of direction;

A more convenient way to compute dot-product is the algebraic formula.

$$\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{d} x_i y_i.$$

Vectors: Illustration of dot product

Consider the five vectors below

$$\mathbf{x} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \ \mathbf{y} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}, \mathbf{p} = \begin{bmatrix} -1 \\ 3 \end{bmatrix}, \ \mathbf{u} = \begin{bmatrix} 3 \\ -1 \end{bmatrix}, \ \mathbf{z} = \begin{bmatrix} -2 \\ -1 \end{bmatrix}$$

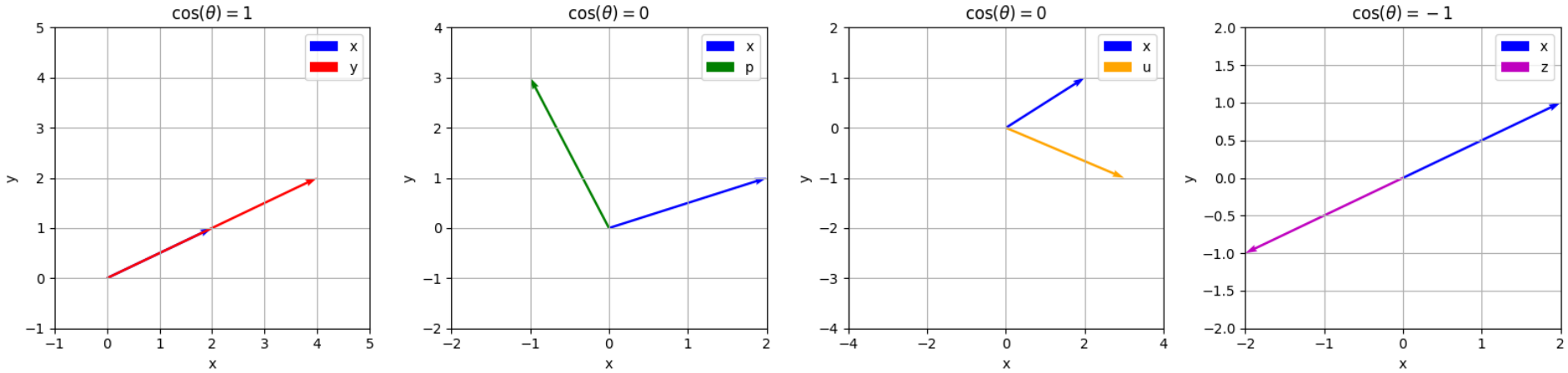

Figure: Illustration of dot product

Vectors: Illustration of dot product

$$\mathbf{x} \cdot \mathbf{y} = \|\mathbf{x}\| \times \|\mathbf{y}\| \times \cos(\theta).$$

$\cos(\theta)$ measures similarity between $\mathbf{x}$ and $\mathbf{y}$ in terms of direction where:

- $\cos(\theta) = 1$ means perfect alignment
- whereas $\cos(\theta) = -1$ means perfect opposition.

Consider the following dot products:

- $\mathbf{x}, \mathbf{y}$ are perfectly aligned resulting into $\cos(\theta) = 1$;
- $\mathbf{x}, \mathbf{p}$ and $\mathbf{x}, \mathbf{u}$ are perpendicular resulting into null dot products because of $\cos(\frac{\pi}{2}) = \cos(\frac{3\pi}{2}) = 0$;
- $\mathbf{x}, \mathbf{z}$ are aligned but look into opposite direction producing $\cos(\theta) = -1$;
- vectors forming angle $\theta \in \left]\frac{3\pi}{2}, \frac{\pi}{2}\right[$ with $\mathbf{x}$ tend to the direction of $\mathbf{x}$;
- inversely, vectors forming angle $\theta \in \left]\frac{\pi}{2}, \frac{3\pi}{2}\right[$ with $\mathbf{x}$ evolve in opposite direction to $\mathbf{x}$.

Matrices

Definition

Visually, matrices tend to be defined as a grid of numbers where elements are identified by row and column indices.

Algebraically, matrices are functions that linearly transform one vector into another. A matrix $\mathbf{W}$ of m rows and d columns (i.e., $\mathbf{W} \in \mathbb{R}^{m \times d}$) is defined as:

$$\mathbf{W} = \begin{bmatrix} w_{11} & \dots & w_{1d} \\ \vdots & \ddots & \vdots \\ w_{m1} & \ddots & w_{md} \end{bmatrix}.$$

$\mathbf{W}$ transforms input vectors $\mathbf{x} \in \mathbb{R}^d$ to output vectors $\mathbf{y} \in \mathbb{R}^m$;

Matrices: Transposition

Definition

Let $\mathbf{A} \in \mathbb{R}^{m \times d}$ be a matrix with m rows and d columns;

$$\mathbf{A} = \begin{bmatrix} a_{11} & \dots & a_{1d} \\ \vdots & \ddots & \vdots \\ a_{m1} & \ddots & a_{md} \end{bmatrix}.$$

Transposed $\mathbf{A}$ noted $\mathbf{A}^T$ is a matrix with d rows and m columns where each j-th column vector in $\mathbf{A}^T$ corresponds to the j-th row vector in $\mathbf{A}$;

$$\mathbf{A}^T = \begin{bmatrix} a_{11} & \dots & a_{m1} \\ \vdots & \ddots & \vdots \\ a_{1d} & \ddots & a_{md} \end{bmatrix}.$$

Matrices: Scalar-Matrix Multiplication

Let $\alpha \in \mathbb{R}$ and $\mathbf{A} \in \mathbb{R}^{m \times d}$, product $\alpha\mathbf{A}$ is defined as

$$\alpha\mathbf{A} = \begin{bmatrix} \alpha a_{11} & \dots & \alpha a_{1d} \\ \vdots & \ddots & \vdots \\ \alpha a_{m1} & \ddots & \alpha a_{md} \end{bmatrix}.$$

$\alpha\mathbf{A} = \mathbf{A}\alpha$

Matrices: Vector-Matrix Multiplication

Let $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{A} \in \mathbb{R}^{m \times d}$, as described above, $\mathbf{Ax}$ is merely the sum

$$\mathbf{Ax} = x_1 \begin{bmatrix} a_{11} \\ \vdots \\ a_{m1} \end{bmatrix} + \cdots + x_d \begin{bmatrix} a_{m1} \\ \vdots \\ a_{md} \end{bmatrix}.$$

The weighted sum above is compactly expressed as sums of products between elements of the vector with elements of matrix rows

$$\mathbf{Ax} = \begin{bmatrix} \sum_{j=1}^{m} a_{1j} x_j \\ \vdots \\ \sum_{j=1}^{m} a_{mj} x_j \end{bmatrix}.$$

Matrices: Matrix-Matrix Multiplication

Product between matrices follows the same principle as matrix-vector product; Consider 2×2 matrices $\mathbf{A}$ and $\mathbf{B}$;

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}.$$

The product $\mathbf{AB}$ is

$$\mathbf{AB} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix},$$

Matrices: Matrix-Matrix Multiplication

Definition

Matrix multiplication can be generalized:
Let $\mathbf{A} \in \mathbb{R}^{m \times d}$ and $\mathbf{B} \in \mathbb{R}^{d \times n}$;

Note that the number of columns in $\mathbf{A}$ must be equal to the number of rows in $\mathbf{B}$;
The dimension of the resulting matrix is $m \times n$

$$\mathbf{A} \times \mathbf{B} = \begin{bmatrix} \sum_{j=1}^{d} a_{1j} b_{1j} & \cdots & \sum_{j=1}^{d} a_{1j} b_{nj} \\ \vdots & \ddots & \vdots \\ \sum_{j=1}^{d} a_{mj} b_{1j} & \cdots & \sum_{j=1}^{d} a_{mj} b_{nj} \end{bmatrix}.$$

Matrices: Few properties on operations between Matrices

$$\mathbf{A}(\mathbf{B}+\mathbf{C}) = \mathbf{AB} + \mathbf{AC}.$$

Multiplication is associative:

$$\mathbf{A}\left(\mathbf{BC}\right) = \left(\mathbf{AB}\right)\mathbf{C}.$$

Multiplication is not commutative:

$$\mathbf{AB} \neq \mathbf{BA}.$$

Transpose of multiplication:

$$\left(\mathbf{AB}\right)^T = \mathbf{B}^T\mathbf{A}^T.$$

Matrices: Special matrices

Matrices are defined by their dimensions but also by the values inside the cells; There are some types of matrices with useful properties:

- Diagonal
- Identity
- Inverse
- Symmetry
- Orthonormality

Diagonal Matrix:
Diagonal matrices are square with null values except in the diagonal

$$\mathbf{D} = \begin{bmatrix} d_{11} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ddots & d_{dd} \end{bmatrix}.$$

Identity Matrix:
Identity matrices are diagonal matrices with 1s in the diagonal;

Identity matrices preserve input vectors both in direction and in length.

Inverse:

Let $\mathbf{A} \in \mathbb{R}^{m \times d}$ be a matrix with m rows and d columns;

We define $\mathbf{A}^{-1} \in \mathbb{R}^d \times \mathbb{R}^m$ as the inverse of $\mathbf{A}$ that is:

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}.$$

Existence of $\mathbf{A}^{-1}$ must satisfy:

- $\mathbf{A}$ must be square (number of rows = number of columns);
- no linear dependence amongst columns of $\mathbf{A}$.

Matrices: Special matrices

Symmetry:
A symmetric matrix is any matrix that is equal to its own transpose:

$$\mathbf{A} = \mathbf{A}^T.$$

Symmetric matrices often arise when entries are generated by some functions of two arguments that do not depend on the order of arguments, such as Covariance or Distance between data.

Orthonormality

An orthogonal matrix is a square whose:

- rows are mutually orthonormal;
  - $A_i \cdot A_j = 0$, for $i \neq j$, where $\cdot$ represents the dot product.
  - $\|A_i\| = 1$, where $\|\cdot\|$ represents the Euclidean norm (magnitude).
- columns are mutually orthonormal.

$$\mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T = \mathbf{I}.$$

The equation above implies that:

$$\mathbf{A}^{-1} = \mathbf{A}^T.$$

Matrix: Determinant

Definition

Matrix determinant $\mathbf{det}(\mathbf{A})$ is a scalar value:

- Whose sign is positive if the order between the basis vectors is unchanged and negative otherwise.
- Whose value equals 0 if at least one column vector of $\mathbf{A}$ is linearly dependent on the others (which leads to space contraction).

For 2×2 matrices

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

determinant is given by $\mathbf{det}(A) = ad - cb$.

Matrix: Determinant

Definition

For 3×3 matrices

$$\mathbf{A} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix},$$

the determinant is computed as follows:

$$\mathbf{det}(\mathbf{A}) = a.\mathbf{det}\left(\begin{bmatrix} e & f \\ h & i \end{bmatrix}\right) - b.\mathbf{det}\left(\begin{bmatrix} d & f \\ g & i \end{bmatrix}\right) + c.\mathbf{det}\left(\begin{bmatrix} d & e \\ g & h \end{bmatrix}\right).$$

The same principle applies to higher dimensions.

Matrices: Eigenvalues & Eigenvectors

Definition

When a matrix $\mathbf{A}$ is applied on input vectors, it tends to alter its length and/or direction;
Under some circumstances, certain vectors $\mathbf{v}$ get altered only in their length, which is

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}, \quad \lambda \neq 0, \mathbf{v} \neq \mathbf{0}.$$

These kinds of vectors are known as eigenvectors, and the coefficient of the length alteration is the corresponding eigenvalue.
Note that if $\mathbf{v}$ is an eigenvector of $\mathbf{A}$, then:

- $\forall\ s \neq 0$, $s\mathbf{v}$ is eigenvector of $\mathbf{A}$;
- $s\mathbf{v}$ still has the same eigenvalue as $\mathbf{v}$.

Matrices: Eigenvalues & Eigenvectors

Definition

Eigenvalues are computed through the characteristic equation defined as

$$\mathbf{Av} = \lambda\mathbf{v}, \mathbf{Av} - \lambda\mathbf{v} = 0,$$

resulting into polynomial in λ where the roots are the eigenvalues of $\mathbf{A}$.

$$\mathbf{det}\left(\mathbf{A} - \lambda\mathbf{I}\right) = 0.$$

Matrix: Matrix factorization - Eigendecomposition

Definition

Factorization splits linear transformations into atomic steps like distinguishing between rotations and length alteration.

Let $\mathbf{A}$ a diagonalizable matrix (square, invertible, and verify other conditions); If $\mathbf{A}$ has m linearly independent eigenvectors $\{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(m)}\}$ with corresponding eigenvalues $\{\lambda_1, \ldots, \lambda_m\}$ then $\mathbf{A}$ can be factorized as;

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}.$$

When m eigenvectors are concatenated into matrix $\mathbf{V}$ and the associated eigenvalues in the diagonal matrix $\mathbf{\Lambda}$, $\mathbf{A}$ can be factorized as

- Where $\mathbf{V}$ is concatenation of m eigenvectors ordered by descending absolute values of eigenvalues.
- If $\mathbf{A}$ is symmetric the matrix of eigenvectors $\mathbf{Q}$ is orthogonal and the factorization becomes

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{T}.$$

Matrix: Matrix factorization - Singular Value Decomposition

Singular Value Decomposition (SVD) provides another way to factorize a matrix;

While eigendecomposition does not apply to any matrix, SVD is;

SVD decomposition has a lot of applications like Image compression, Recommender Systems (Netflix, Spotify, Amazon, etc.);

Matrix: Matrix factorization - Singular Value Decomposition

Definition

Let $\mathbf{M}$ be a $m \times d$ matrix. The SVD decomposition of $\mathbf{M}$ is

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T,$$

where for a given input vector $\mathbf{x}$:

- $\mathbf{V}^T$ is $d \times d$ orthogonal matrix and performs rotation on $\mathbf{x}$ (we call $\mathbf{x}_V$);
- $\mathbf{\Sigma}$ is $m \times d$ and performs stretching of $\mathbf{x_V}$ then projects it into m dimensions (we call $\mathbf{x}_{V\Sigma}$);
- $\mathbf{U}$ is $m \times m$ orthogonal matrix and performs last rotation on $\mathbf{x}_{V\Sigma}$.

Matrix: Matrix factorization - Singular Value Decomposition

Definition

$\mathbf{V}$ is constructed by the eigenvectors of $\mathbf{M}^T\mathbf{M}$;

$\mathbf{U}$ is constructed by the eigenvectors of $\mathbf{M}\mathbf{M}^T$;

$\mathbf{\Sigma}$ is a diagonal matrix whose entries are called singular values and correspond to the square root of the eigenvalues of $\mathbf{M}\mathbf{M}^T$ or $\mathbf{M}^T\mathbf{M}$;

The number of non-zero singular values equals the rank of $\mathbf{M}$;

If $\mathbf{\Sigma}$ is not square, the rest of the diagonal is filled with zeros.

Matrix: python illustration

```
import numpy as np
#Creation of a 2 X 2 matrix
A = np.array([[2,1],[1,2]], dtype = np.float32)
print('Matrix A has the dimension ' + str(A.shape))
print(A)
#Product between two matrices
B = np.array([[3,4,1], [4,7,0]])
AB = A.dot(B)
print('Dimension of A.B = ' + str(AB.shape))
print(AB)
#Transpose
ABt = AB.transpose()
print('Dimension of transposed AB = ' + str(ABt.shape))
print('\nTransposed AB = ')
print(ABt)
#Diagonal Matrix
D = np.diag([1,2,3,4,5,6])
print(D)
```

Matrix: python illustration

```python
import numpy as np

#Determinant
A = np.array([[1, 2], [3, 4]])
detA = np.linalg.det(A)
print('Determinant of A = ' + str(detA))

#eigenvalues & eigenvectors
A = np.array([[2,1],[1,2]], dtype = np.float32)
lambdas, V = np.linalg.eig(A)
print('eigenvalues = ')
print(lambdas)
print('eigenvectors = ')
print(V)

#SVD decomposition
U, Sigma, Vt = np.linalg.svd(A)
```

Outline

Why probability ?

- A key concept in the field of machine learning and data science is that of uncertainty.
- The uncertainty arises through the noise on measurements.
- Probability theory provides a consistent framework for the quantification and manipulation of uncertainty.
- It allows us to make optimal predictions given all the available information, even though it may be incomplete.

Random experiment

Definition

The term "random experiment" refers to a renewable experiment that does not necessarily yield the same result when repeated under identical conditions.

Example: Imagine randomly picking one of the boxes, randomly selecting a ball, and repeating the process an infinite number of times.

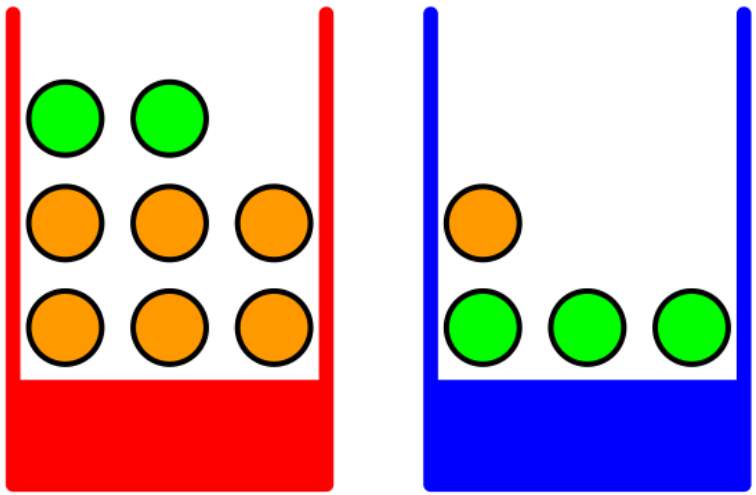

Figure: Two colored boxes, each containing green and orange balls

Sample space and event

Definition

The set of all possible outcomes of a given random experiment is called the sample space or state space and is denoted Ω. Therefore, an element of the sample space is an outcome of the experiment. A subset of Ω is called an event.

Example: In the experiment of choosing the box, the sample space is $\Omega = \{B, R\}$, where R (respectively B) is the fact of choosing the red (respectively blue) box.

σ-algebra

Definition

A family $\mathcal{A}$ of subsets of the sample space Ω is called a "σ-algebra" if it satisfies the following three properties:

1. $\Omega \in \mathcal{A}$.
2. If $A \in \mathcal{A}$ then its complement $A^c \in \mathcal{A}$.
3. If $A, B \in \mathcal{A}$ then their union $A \cup B \in \mathcal{A}$.

The elements of $\mathcal{A}$ are events.

Probability

Definition

A probability (or probability measure) on $(\Omega, \mathcal{A})$ is a function:

$$\mathbb{P} : \mathcal{A} \longrightarrow [0, 1]$$

That satisfies the following three Kolmogorov axioms:

- $0 \leq \mathbb{P}(A) \leq 1$ for all event $A \in \mathcal{A}$.
- $P(\Omega) = 1$.
- If $A_1, A_2, A_3, \ldots$ are mutually exclusive (disjoint) events in $\mathcal{A}$, then:

$$\mathbb{P}(A_1 \cup A_2 \cup A_3 \cup \ldots) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) + \ldots$$

Example: Imagine we choose the red box 40% of the time, and we pick the blue box 60% of the time. We denote by the R the event of choosing the red box. Then the probability of selecting the blue box $\mathbb{P}(R)$ is 0.4.

Proposition

- $\mathbb{P}(A) = 1$ (respectively, $\mathbb{P}(A) = 0$) means that the event A will definitely (respectively, never) happen. In particular, we have $\mathbb{P}(\Omega) = 1$ and $\mathbb{P}(\emptyset) = 0$.
- $\bar{A}$ is the event not A, and $\mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A)$.
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - P(A \cap B) \leq P(A) + P(B)$.
- If $A \subset B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

Example: $\bar{R} = B$ is the event of choosing the blue box, and $\mathbb{P}(B) = 0.6$.

Uniform distribution

Definition

Let Ω be a finite discrete sample state. A discrete uniform distribution is a probability that associates to each element $\omega \in \Omega$ the same value (same probability of occurring). More formally, if $|\Omega|$ denotes the cardinal of Ω, then

$$\forall \omega \in \Omega,\ \mathbb{P}(\omega) = \frac{1}{|\Omega|}$$

Example:

- Consider the throwing dice experience, then $\Omega = \{1, 2, 3, 4, 5, 6\}$ and:

$$\mathbb{P}(1) = \mathbb{P}(2) = \mathbb{P}(3) = \mathbb{P}(4) = \mathbb{P}(5) = \mathbb{P}(6) = \frac{1}{6}$$

- Consider the flipping coin experience, where "0" denotes heads and "1" tails. Then we have $\Omega = \{0, 1\}$, and

$$\mathbb{P}(0) = \mathbb{P}(1) = \frac{1}{2}$$

- Is the probability of choosing the blue or red box uniform? Why?

Uniform distribution

Proposition

For all events A (subset of Ω) we have

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}$$

Example: Imagine throwing two dice. We compute the probability of the event A "the sum of the two results will equal 7"

We have $\Omega = \{1, \cdots, 6\} \times \{1, \cdots, 6\} = \{(1,1), \cdots, (6,6)\}$ and $A = \{(1,6), (6,1), (5,2), (2,5), (4,3), (3,4)\}$. Hence

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{6}{36} = \frac{1}{6}$$

Independence

Definition

Two events A and B are independent if and only if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

Example: Imagine throwing two dice. We denote by A the event of obtaining 5 with the first dice and B the event of obtaining 3 with the second dice. Are A and B independent?

$$A = \{(5, k), k = 1, \cdots, 6\},\ B = \{(k, 3), k = 1, \cdots, 6\} \text{ and } A \cap B = \{(5, 3)\}$$

Hence,

$$\mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(A \cap B) = \frac{1}{36}$$

Conditional probability

Definition

Let $A, B \in \mathcal{A}$ two events. The conditional probability of A given B is defined by:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

This is defined only if $\mathbb{P}(B) > 0$, it is clear that if $\mathbb{P}(B) = 0$, then $\mathbb{P}(A|B) = \mathbb{P}(A)$.

Independence and conditional probability

Proposition

Let $A, B \in \mathcal{A}$ two events, then:

1. $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$.
2. $\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B))}$.

Theorem

The events A and B are independent if and only if $\mathbb{P}(A|B) = \mathbb{P}(A)$.

Theorem (Chain rule)

Let $A_1, \ldots, A_n$ be events, then we have:

$$P(A_1, \cdots, A_n) = P(A_1) \prod_{i=2}^{n} P(A_i|A_1, \cdots, A_{i-1})$$

Bayes rule

Theorem

Let A be an event, and $A_1, \ldots, A_n$ events mutually disjoints, such that for every i, $\mathbb{P}(A_i) > 0$, and $\mathbb{P}(A_1, \sqcup, \ldots \sqcup A_n) = 1$, then we have for every i

$$\mathbb{P}(A_i|A) = \frac{\mathbb{P}(A|A_i)\mathbb{P}(A_i)}{\sum_{j=1}^{n} \mathbb{P}(A|A_j)\mathbb{P}(A_j)}$$

Example:

B and R denote the events of choosing the blue and red boxes, respectively. We also denote the events of choosing the green and orange balls by G and O. We have

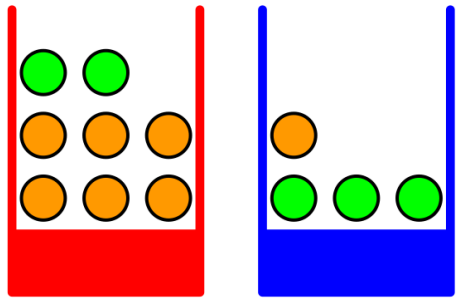

Figure: The example of choosing a box and choosing a ball.

Bayes rule

We have

- $\mathbb{P}(B) = 3/5$ and $\mathbb{P}(R) = 2/5$.
- $\mathbb{P}(G|B) = 3/4$ and $\mathbb{P}(O|B) = 1/4$.
- $\mathbb{P}(G|R) = 1/4$ and $\mathbb{P}(O|R) = 3/4$.
- $\mathbb{P}(G) = \mathbb{P}((G \cap R) \sqcup (G \cap B)) = \mathbb{P}(G \cap R) + \mathbb{P}(G \cap B) = \mathbb{P}(G|B)\mathbb{P}(B) + \mathbb{P}(G|R)\mathbb{P}(R) = \frac{11}{20}$.
- $\mathbb{P}(O) = 1 - \mathbb{P}(G) = 1 - \frac{11}{20} = \frac{9}{20}$.
- $\mathbb{P}(B|G) = \frac{P(G|B)P(B)}{P(G|B)P(B)+P(G|R)P(R)} = \frac{1}{3}$
- $\mathbb{P}(R|G) = \frac{2}{3}$

$\mathbb{P}(B)$ and $\mathbb{P}(R)$ can be viewed as prior probabilities, $\mathbb{P}(B|G)$ and $\mathbb{P}(R|G)$ can be viewed as posterior probabilities.

Random variables

Definition

A random variable is a variable that can take on different values randomly. A random variable is a function X that goes from the sample of states Ω to a set E.

Remark

Let $x \in E$ and $I \subset E$, then the subset $\{\omega \in \Omega | X(\omega) = x\}$ is an event. Hence it has the sense to compute $\mathbb{P}(\{\omega \in \Omega | X(\omega) = x\})$ and $\mathbb{P}(\{\omega \in \Omega | X(\omega) \in I\})$. In the following, we will denote $\mathbb{P}(X = x)$ to design $\mathbb{P}(\{\omega \in \Omega | X(\omega) = x\})$ and $\mathbb{P}(X \in I)$ to design $\mathbb{P}(\{\omega \in \Omega | X(\omega) \in I\})$.

Cumulative distribution function

Definition

Let X be a real random variable(i.e., $E \subset \mathbb{R}$), the cumulative distribution function of X, denoted by F_X is defined by:

$$F_X(t) = \mathbb{P}(X \leq t)$$

Proposition

F_X satisfies the following properties

- F_X is a non-decreasing function.
- $\lim_{t \to -\infty} F_X(t) = 0$.
- $\lim_{t \to +\infty} F_X(t) = 1$.

Probability distributions

Definition

Let X be a random variable. The probability distribution of X is given by the set $\{\mathbb{P}(X = x), x \in E\}$. A probability distribution describes how likely a random variable is to take on each of its possible states. In the following, we will denote $p()$, the probability distribution of X, instead of $\mathbb{P}()$.

Discrete variables

Definition

A discrete random variable has a finite or countably infinite number of states (i.e., E is discrete). Note that these states are not necessarily integers; they can also just be named states that are not considered to have any numerical value.

Example:

1. Let be X, the random variable that describes the chosen box (red or blue). Then, we have:

$$p(X = r) = 0.4 \text{ and } p(X = b) = 0.6$$

2. Let be X the random variable that returns the result of throwing dice. Then X takes its values in $\{1, 2, 3, 4, 5, 6\}$, which is a discrete set, and we have $\forall k \in \{1, 2, 3, 4, 5, 6\}$:

$$p(X = k) = \frac{1}{6}$$

Discrete variables: Properties

Proposition

Let X and Y be two discrete variables

- Sum rule: $\sum_x p(X = x)) = 1$
- Joint probability distribution: $p(X = x, Y = y)$ denotes the probability that $X = x$ and $Y = y$ simultaneously. We may also write $p(x, y)$ for brevity.
- Marginal distribution: $p(X = x) = \sum_y p(X = x, Y = y)$
- Conditional distribution: $p(X = x|Y = y) = \frac{p(X=x,Y=y)}{P(Y=y)}$
- Bayes rule:

$$p(X = x|Y = y) = \frac{p(Y = y|X = x)p(X = x)}{\sum_{x'} P(Y = y|X = x')P(X = x')}$$

- Chain rule: $P(X_1, \cdots, X_n) = P(X_1) \prod_{i=2}^{n} P(X_i|X_1, \cdots, X_{i-1})$

Discrete variables: Expectation

Definition

Let X be a (real) discrete variable. The expectation of X is defined by:

$$\mathbb{E}[X] = \sum_{x \in E} x \cdot p(X = x)$$

Proposition

Let X and Y be two (real) discrete variables. $\lambda \in \mathbb{R}$

- $\mathbb{E}[\lambda] = \lambda$.
- $\mathbb{E}[X + \lambda Y] = \mathbb{E}[X] + \lambda \mathbb{E}[Y]$.
- If $X \geq 0$ then $\mathbb{E}[X] \geq 0$.
- If $X \geq Y$ then $\mathbb{E}[X] \geq \mathbb{E}[Y]$.

Discrete variables: Variance

Definition

Let X be a (real) discrete variable. The variance of X is defined by:

$$\mathbb{V}[X] = \mathbb{E}\left[(X - E[X])^2\right]$$

Proposition

Let X be a (real) discrete variable. $\lambda, \beta \in \mathbb{R}$

- $\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.
- $\mathbb{V}[X] \geq 0$.
- $\mathbb{V}[\lambda X + \beta] = \lambda^2 \mathbb{V}[X]$.
- If $\mathbb{V}[X] = 0$ X is a constant variable.

Definition

Let X be a (real) discrete variable. The standard deviation of X is defined by:

$$\sigma(X) = \sqrt{\mathbb{V}[X]}$$

Independent random variables

Definition

Let $X_1, \cdots X_n$ be random variables that take their values in $E_1, \cdots E_n$, respectively. We say that the variables $X_1, \cdots X_n$ are independent if and only if $\forall x_1 \in E_1, \cdots, \forall x_n \in E_n$

$$\mathbb{P}(X_1 = x_1, \cdots, X_n = x_n) = \mathbb{P}(X_1 = x_1) \times \cdots \times \mathbb{P}(X_n = x_n)$$

Proposition

Let X and Y be two (real) independent random variables.

- $\mathbb{E}[XY] = \mathbb{E}[X] \times \mathbb{E}[Y]$
- $\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y]$

Common probability distributions: Bernoulli

The Bernoulli distribution models the experiences with two issues (success and failure, heads and tails) represented by 1 and 0.

Definition

A random variable X follows the Bernoulli distribution with parameter p (denoted by $X \sim \mathcal{B}(p)$) if X takes only two values 0 and 1 and

$$\mathbb{P}(X = 1) = p \text{ and } \mathbb{P}(X = 0) = 1 - p$$

Proposition

If $X \sim \mathcal{B}(p)$ then

- Expectation: $\mathbb{E}[X] = p$
- Variance: $\mathbb{V}[X] = p(1 - p)$

Example: Let X be the results of the flipping coin experience. We have

$$\mathbb{P}(X = 1) = \mathbb{P}(X = 0) = 0.5$$

Then $X \sim \mathcal{B}(0.5)$

Common probability distributions: Binomial

Assume we repeat n times the experience of Bernoulli with parameter p. The random variable X that follows the Binomial distribution returns the number of successes of the n experiences.

Definition

A random variable X follows the Binomial distribution with parameters $p \in [0,1]$ and $n \in \mathbb{N}$ (denoted by $X \sim \mathcal{B}(n,p)$) if X takes its values $\{0, \cdots, n\}$ and

$$\forall k \in \{1, \cdots, n\},\ \mathbb{P}(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

Proposition

If $X \sim \mathcal{B}(n,p)$ then

- Expectation: $\mathbb{E}[X] = np$
- Variance: $\mathbb{V}[X] = np(1-p)$

Common probability distributions: Geometric

Assume we independently repeat the experience of Bernoulli with parameter $p \in]0, 1]$. The random variable X that follows the Geometric distribution returns the rank of the first success, i.e., $\mathbb{P}(X = k)$ is the probability that the first success will occur at k-th iteration.

Definition

A random variable X follows the Geometric distribution with parameter $p \in]0, 1]$ (denoted by $X \sim \mathcal{G}(p)$) if X takes its values $\mathbb{N}^*$ and

$$\forall k \in \mathbb{N}^*, \ \mathbb{P}(X = k) = p \cdot (1 - p)^{k-1}$$

Proposition

If $X \sim \mathcal{G}(p)$ then

- Expectation: $\mathbb{E}[X] = \frac{1}{p}$
- Variance: $\mathbb{V}[X] = \frac{1-p}{p^2}$

Common probability distributions: Poisson

Consider an event that occurs on average λ times during a given time interval. A random variable X, following the Poisson distribution with parameter λ, will give the number of times the event occurs during that time interval.

Definition

A random variable X follows the Poisson distribution with parameter $\lambda \in \mathbf{R}^*$ (denoted by $X \sim \mathcal{P}(\lambda)$) if X takes its values $\mathbb{N}$ and

$$\forall k \in \mathbb{N},\ \mathbb{P}(X = k) = \frac{\lambda^k}{k!} \times e^{-\lambda}$$

Proposition

If $X \sim \mathcal{G}(p)$ then

- Expectation: $\mathbb{E}[X] = \lambda$
- Variance: $\mathbb{V}[X] = \lambda$

Continuous variables

Definition

A probability density function is a positive function f that satisfies $\int_{\mathbb{R}} f(x)dx = 1$

Definition

Let Ω be a state space, f a probability density function, and $X : \Omega \longrightarrow \mathbb{R}$ a random variable. f is the density function of f if for all $a, b \in \mathbb{R}$

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)dx$$

Cumulative distribution function

Definition

Let X be a continuous random variable with a density f. The cumulative distribution function (cdf) associated to X is defined by

$$F_X(t) = \mathbb{P}(X \leq t), \ t \in \mathbb{R}$$

Proposition

Let X be a continuous random variable with F_X it cdf, $a, b \in \mathbb{R}$, such that $a < b$. Hence,

$$\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a)$$

Proposition

Two random variables have the same probability distribution if they have the same cdf.

Expectation and variance

Definition

Let X be a continuous random variable with a density f. The expectation of X is defined by:

$$\mathbb{E}[X] = \int_{x \in \mathbb{R}} x \cdot f(x)$$

Definition

Let X be a continuous random variable. The variance of X is defined by:

$$\mathbb{V}[X] = \mathbb{E}\left[(X - E[X])^2\right]$$

Common probability distributions: Uniform

Definition

Let $[a, b] \subset \mathbb{R}$, the uniform distribution (denoted by $\mathcal{U}([a, b])$) is defined by the following density function

$$f(x) = \frac{1}{b-a} \mathbb{1}[a, b](x), \ x \in \mathbb{R}$$

Proposition

If $X \sim \mathcal{U}([a, b])$

- Expectation: $\mathbb{E}[X] = \frac{a+b}{2}$
- Variance: $\mathbb{V}[X] = \frac{(b-a)^2}{12}$

Common probability distributions: Exponential

Definition

Let $\lambda \subset \mathbb{R}_+^*$, the exponential distribution (denoted by $\mathcal{E}(\lambda)$) is defined by the following density function

$$f(x) = \lambda e^{-\lambda x}, \ x \in \mathbb{R}_+$$

Proposition

If $X \sim \mathcal{E}(\lambda)$

- Expectation: $\mathbb{E}[X] = \frac{1}{\lambda}$
- Variance: $\mathbb{V}[X] = \frac{1}{\lambda^2}$

Common probability distributions: Normal distribution

Definition

Let $\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+^*$, the normal (Gaussian) distribution (denoted by $\mathcal{N}(\mu, \sigma^2)$) is defined by the following density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{-(x-\mu)^2}{2\sigma^2}}, \ x \in \mathbb{R}$$

Proposition

If $X \sim \mathcal{N}(\mu, \sigma^2)$

- Expectation: $\mathbb{E}[X] = \mu$
- Variance: $\mathbb{V}[X] = \sigma^2$

Outline

Descriptive statistics

Descriptive statistics

- The main objective of descriptive statistics is to extract relevant and significant information from a set of observed data $(x_1, \cdots, x_n)$.
- The set $(x_1, \cdots, x_n)$ is called a sample, data, observed data, or dataset.
- n represents the sample size.
- These observed data may come from surveys (sociology, psychology, etc.), scientific experiments (medical, biology, etc.), or historical records (such as meteorological, finance, etc.).
- These data might be voluminous and complex, which makes them challenging to interpret and exploit. Hence, we need to summarize them and find relevant tools to visualize them.

Descriptive statistics

- The descriptive statistics approach introduces a probabilistic model that is the most susceptible to generating the data.
- The observed data $x_1, \cdots, x_n$ are assumed to be the realization of some random variables $X_1, \cdots, X_n$.
- Descriptive statistics seeks to answer this kind of question:
  1. Are these variables independent?
  2. Do they have the same law?
  3. What is the shape of the law?
  4. What are the parameters of this law?

Univariate analysis: Variables

We suppose that the observations $x_1, \cdots, x_n$ are real-valued. We assume that these variables are the realization of random variables $X_1, \cdots, X_n$.

In the following, we assume that these variables are independent and identically distributed (i.i.d), which means that these variables are independent and follow the same law F.

We distinguish two main types of data: data that came from **continuous variables** and those that came from **discrete variables**.

Univariate analysis: continuous vs. discrete variables

We say that the real variables $X_1, \cdots, X_n$ are continuous if and only if the law F is continuous for example of real variables: Speed, height, temperature, blood pressure, etc.

Discrete variables can take a finite number of possible values (less than the sample size n). We distinguish two types of discrete variables:

Quantitative variables such as counts: number of children per family, number of students in a class, etc.

Qualitative variables, also called categorical variables. They represent modalities, categories, or even levels. Qualitative variables are divided into two types: nominal (no ordering), for example, gender (male or female), colors, etc. Ordinal (order implied) such as satisfaction rating ("extremely dislike", "dislike", "neutral", "like", "extremely like").

Note that the algebraic operations (such as addition, multiplication) have no sense for qualitative variables.

Univariate analysis: continuous vs. discrete variables

```
import pandas as pd # import pandas library
df = pd.read_csv("Medical_insurance.csv") # Load csv file
df.head(5) # display first 15 rows of the dataframe
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Figure: Medical Insurance Price: first five individuals

Graphical representation: edf

Definition

The empirical distribution function (edf) associated to the sample $x_1, \cdots, x_n$ is defined by:

$$\hat{F}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{x_i \leq t\}} = \frac{|\{i,\, x_i \leq t\}|}{n},\ t \in \mathbb{R}$$

- This function is non-decreasing and takes its values in $[0, 1]$.
- It verifies

$$\hat{F}(t) = 0,\ \forall t < \min\{x_1, \cdots, x_n\} \text{ and } \hat{F}(t) = 1,\ \forall t > \max\{x_1, \cdots, x_n\}$$

- When $x_1, \cdots, x_n$ are i.i.d and follow the same law F, then the empirical function $\hat{F}$ approximates F.
- $\hat{F}$ is called the empirical law associated to $x_1, \cdots, x_n$.

Graphical representation: edf

```
import matplotlib.pyplot as plt
from statsmodels.distributions.empirical_distribution import ECDF

# Compute the empirical (cumulative) distribution function
ecdf_children = ECDF(list(df["children"][0:50]))
ecdf_bmi = ECDF(list(df["bmi"][0:50]))

# Plot the two empirical functions
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(15,6))
plt.rcParams['axes.spines.right'] = False
plt.rcParams['axes.spines.top'] = False

ax1.scatter(ecdf_children.x,ecdf_children.y,color="black", s=50, marker='.')
ax1.set_axisbelow(True)
ax1.grid()
ax2.scatter(ecdf_bmi.x,ecdf_bmi.y, color="black", s=50, marker='.')
ax2.set_axisbelow(True)
ax2.grid()
plt.show()
```

Graphical representation: edf

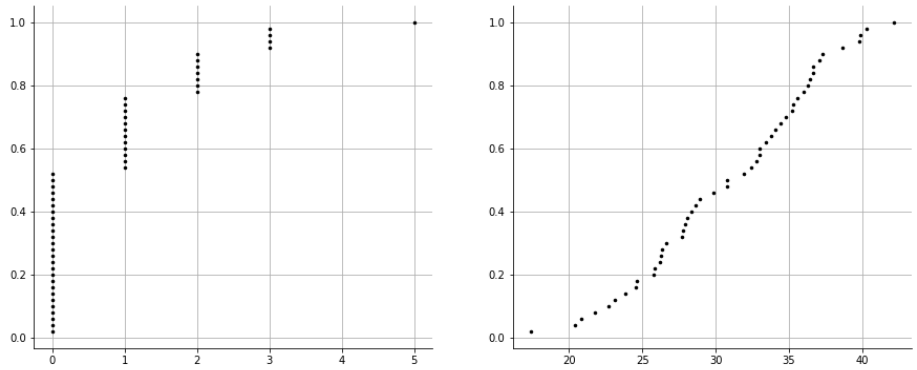

Figure: Edf for the number of children (left) and BMI (right) with $n = 50$

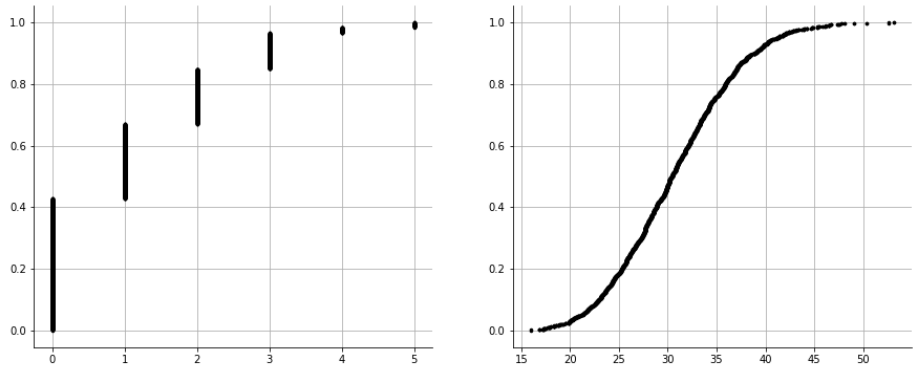

Figure: Edf for the number of children (left) and BMI (right) with $n = 2772$

Graphical representations: Bar plot

Definition

The bar plot can be used to visualize discrete data. We denote by $\mathcal{M} = \{m_1, \cdots, m_K\}$ the set of observed modalities with $K \leq n$. For each $k \in \{1, \cdots, K\}$ we compute the proportion $\hat{p}_k$ of observations with the modality m_k in the sample defined by

$$\hat{p}_k = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i = m_k\}} = \frac{|i, X_i = m_k|}{n}, \ k = 1, \cdots, K.$$

- To draw a bar graph, we draw vertical bars with the modalities in the $x-$axis with proportions $\hat{p}_k$ in $y-$axis.
- The vector $(\hat{p}_1, \cdots, \hat{p}_K)$ defines a probability law over the set of modalities $\mathcal{M}$. In fact, we have

$$\forall \{1, \cdots, K\}, \ \hat{p}_k \in [0, 1] \text{ and } \sum_{k=1}^{K} \hat{p}_k = 1,$$

and $\hat{p}_k \approx \mathbb{P}(X = m_k)$.

Graphical representations: Bar plot

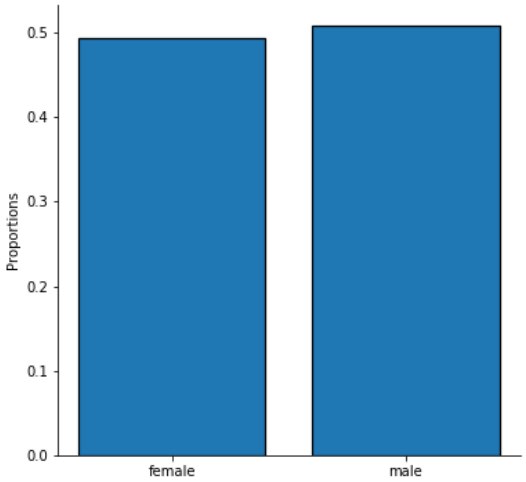

Figure: Bar plot of males and females

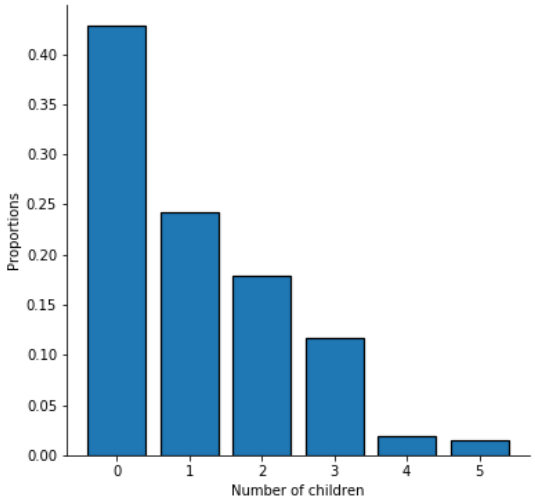

Figure: Bar plot of the number of children

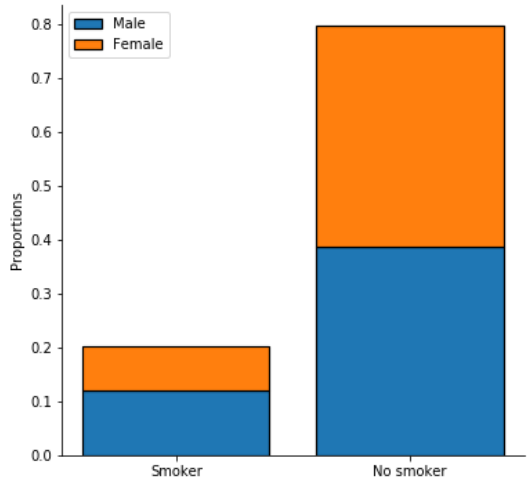

Figure: Bar plot of smokers and non-smokers for males and females

Graphical representations: Histogram

Definition

The histogram can be used to visualize data that came from continuous variables. We let $I = [a, b]$ be the interval that contains all the observations, K the number of partitions of I in K intervals $I_k = [a + (k-1)h, a + kh]$, note that all the intervals I_k have the same length $h = (b-a)/K$. Finally, we let n_k be the number of observations x_i that belong to the interval I_k:

$$n_k = |\{i, x_i \in I_k\}| = \sum_{i=1}^{n} \mathbb{1}_{\{x_i \in I_k\}}$$

Hence, the histogram is a function defined by

$$\hat{f}(x) \frac{1}{nh} \sum_{k=1}^{K} n_k \mathbb{1}_{\{x \in I_j\}}, \; x \in \mathbb{R}$$

Graphical representations: Histogram

- The shape of the histogram depends on the choice of the width h of the intervals A_k.
- The choice of the h depends on the shape of the density f and the sample size n.
- The area $S_k = h\hat{f}(x), x \in I_k$ of the rectangle associated with the interval I_k estimates $\mathbb{P}(X \in A_k)$ the probability that an observation belongs to the interval I_k.
- The histogram $\hat{f}$ is an estimator of the density f.

Graphical representations: Histogram

```
# Plot histogram for BMI variable by setting number bins of bins to 15
fig, ax = plt.subplots(figsize=(7,6))
ax.hist(df["bmi"], bins = 15, edgecolor='black', linewidth=1.2, density = True)
plt.grid()
ax.set_axisbelow(True)
ax.set_ylabel("Frequency")
ax.set_xlabel("BMI")
plt.show()
```

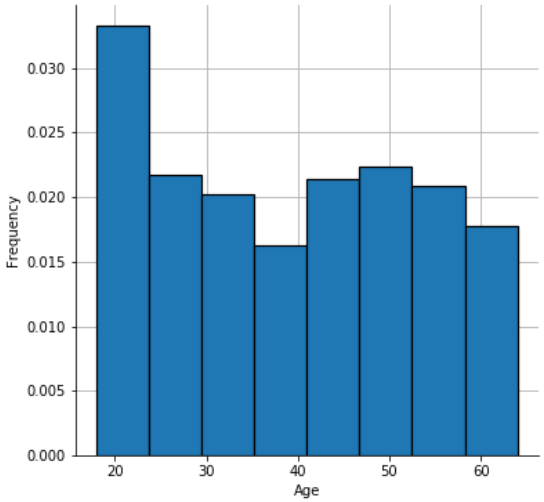

Figure: Histogram of Age

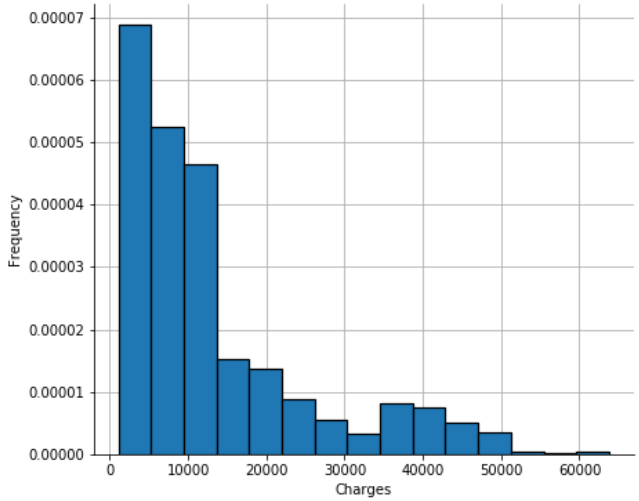

Figure: Histogram of Charges

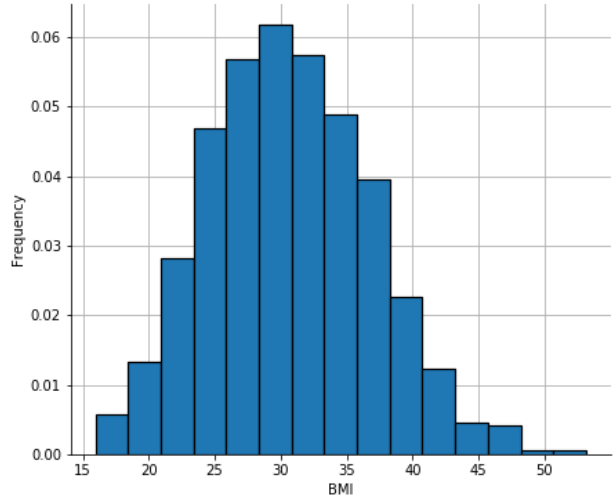

Figure: Histogram of BMI

Graphical representations: Histogram

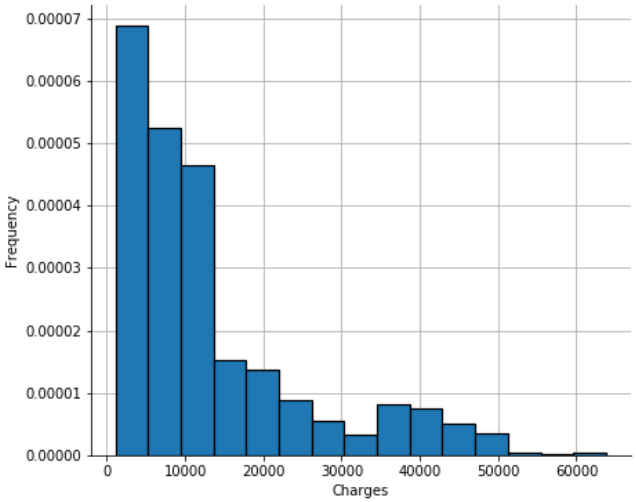

Figure: Bins = 15

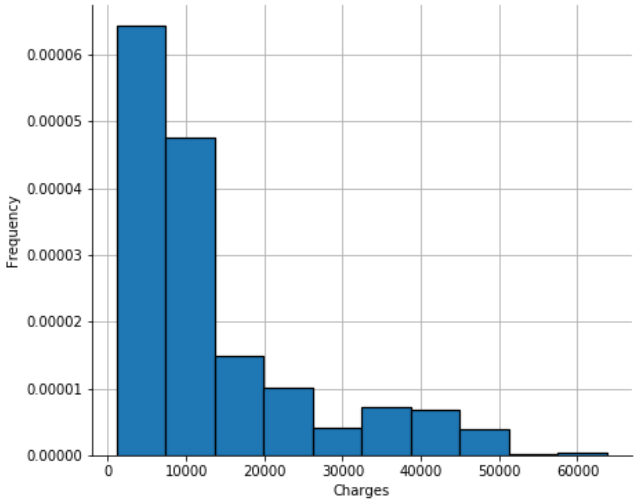

Figure: Bins = 10

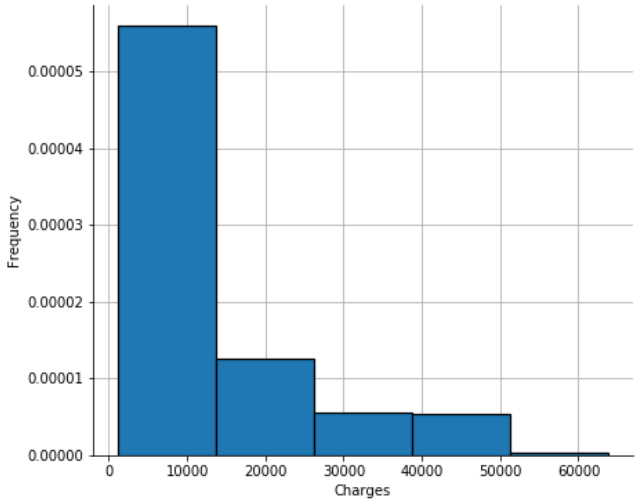

Figure: Bins = 5

Definition

Order statistics $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ associated to the sample $X_1, \cdots, X_n$ are the values X_i placed in ascending order:

$$X_j \in \{X_1, \cdots, X_n\},\ X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

The k-th order statistic of the sample is its $k-$th smallest value.

In particular, we have:

$$X_{(1)} = \min\{X_1, \cdots, X_n\} \text{ and } X_{(n)} = \max\{X_1, \cdots, X_n\}$$

Univariate analysis: Statistical indicators

Definition

The empirical quantile $x_\alpha(n)$ of order $\alpha \in]0,1[$ associated to $\{x_1, \cdots, x_n\}$ is defined

$$x_\alpha(n) = \hat{F}^{-1}(\alpha) = \inf\{t \in \mathbb{R},\ \hat{F}(t) \geq \alpha\}$$

where $\hat{F}^{-1}$ is generalized inverse of the empirical distribution function $\hat{F}$.

- The empirical quantile $x_\alpha(n)$ can be expressed using order statistics. In fact, we have $\forall \alpha \in]0,1[$:

  $$x_\alpha(n) = X_{(\lceil \alpha n \rceil)},$$

  where $\lceil t \rceil$ denotes the upper integer part of t, which is the smallest integer greater than or equal to t (i.e. $\lceil t \rceil = \min\{p \in \mathbb{Z}, p \geq t\}$)
- The empirical quantile $x_\alpha(n)$ of order α is a value that slices the sample in two portions such that $\alpha 100\%$ of the observations are lower than $x_\alpha(n)$ and $(1-\alpha)100\%$ of the observations are greater than $x_\alpha(n)$. We have:

  $$|\{i, X_i \leq x_\alpha(n)\}| \geq \alpha n \quad \text{and} \quad |\{i, X_i \geq x_\alpha(n)\}| \geq (1-\alpha)n$$

- $x_{1/2}(n)$ is the empircal median, $x_{1/4}(n)$ and $x_{3/4}(n)$ are the first and third quartiles.

The central tendency

The central tendency refers to the point or value around which most of the observations are concentrated. Two main measures of the central tendency are the empirical mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

and the empirical median $x_{1/2}(n)$.
In general, the empirical mean differs from the empirical mean.

Graphical representation: Box plot

```
import seaborn as sns

# Box plot
b = sns.boxplot(data = df, y = "charges")
b.get_figure();
```

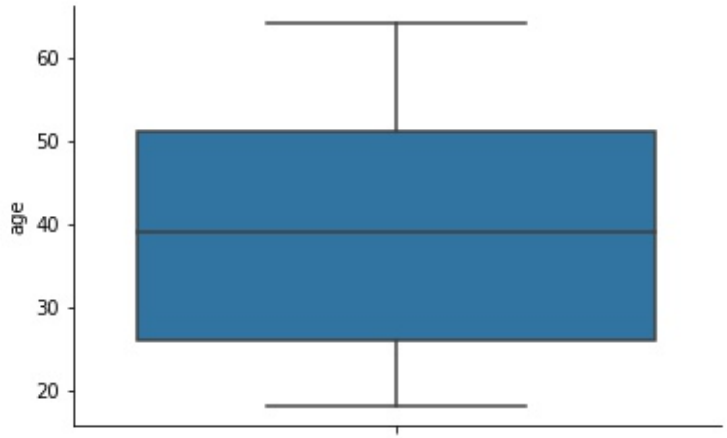

Figure: Box plot of different ages

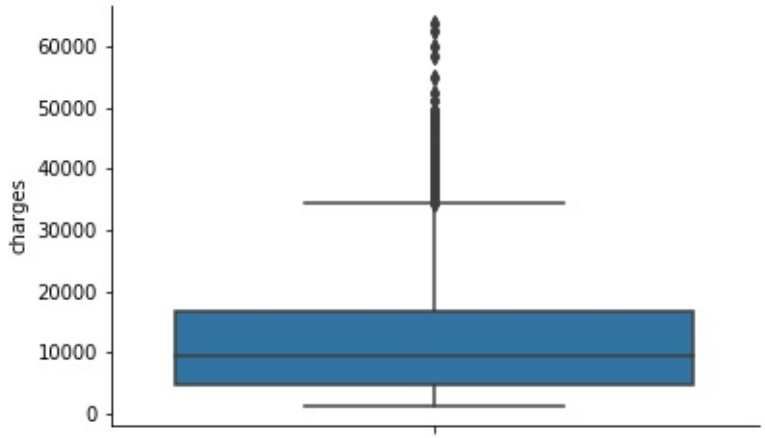

Figure: Box plot of charges

Graphical representation: Box plot

```
# Add the points
b = sns.stripplot(data = df,
                  y = "age",
                  color = "crimson",
                  linewidth = 1,
                  alpha = 0.4)
```

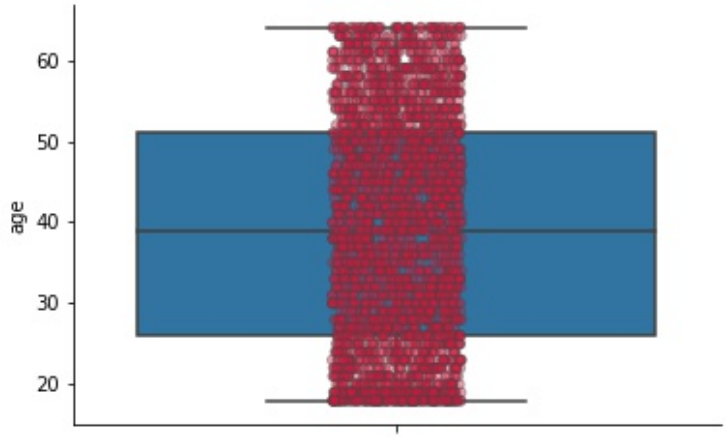

Figure: Box plot of different ages

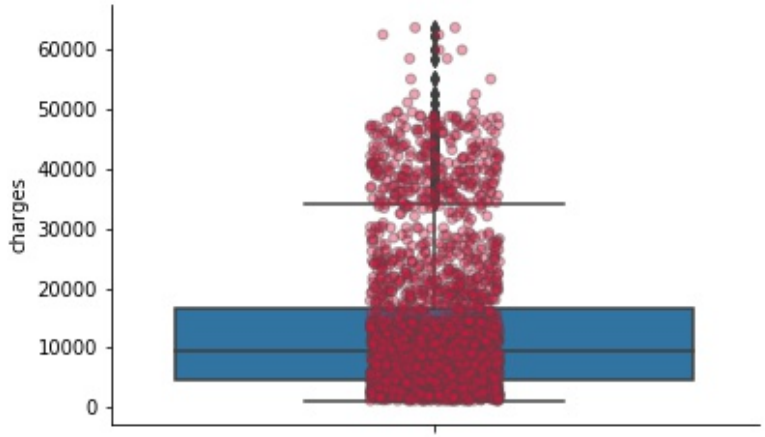

Figure: Box plot of charges

Graphical representation: Box plot

```
# Box plot for charges by regions
b = sns.boxplot(data = df, x = "region", y = "charges")
b.get_figure();
```

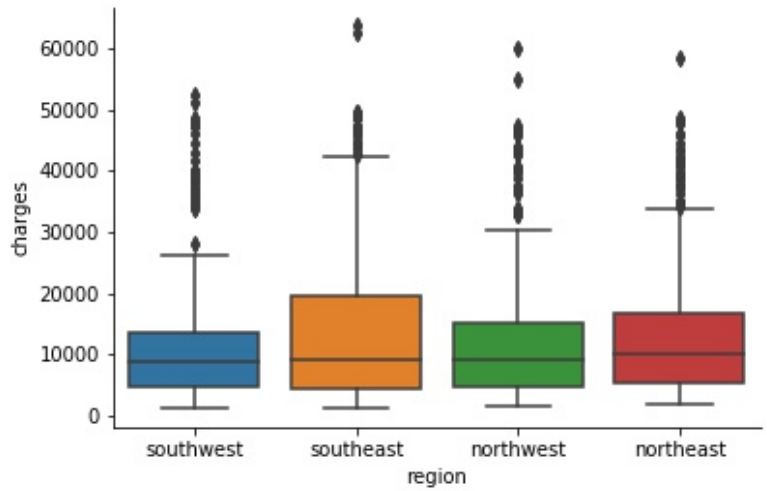

Figure: Box plot of charges by regions

Graphical representation: Box plot

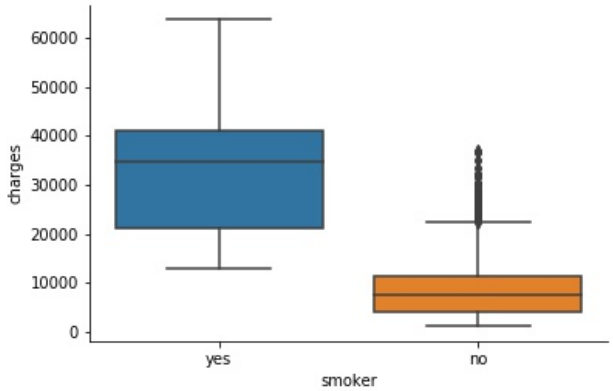

Figure: Box plot of charges for persons who smoke and those who don't

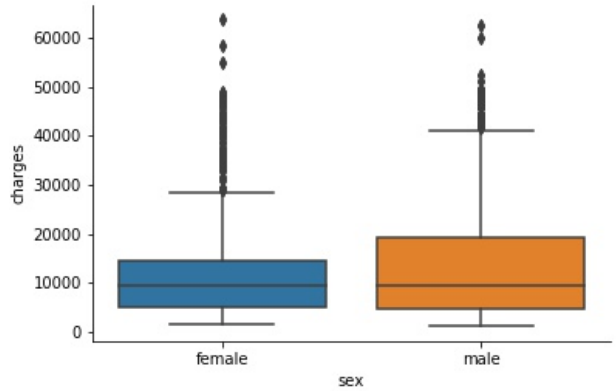

Figure: Box plot of charges by gender

Univariate analysis: Statistical indicators

Dispersion

In the context of random variables $X_1, X_2, \ldots, X_n$, dispersion refers to the extent of variability or spread among these variables. It quantifies how the individual random variables deviate from a central tendency measure (such as the mean or median) and provides information about the data distribution.

Dispersion measures

There are several commonly used measures of dispersion for random variables:

- Range: The range is the difference between the maximum and minimum values of the random variables.

$$\text{Range} = X_{(n)} - X_{(1)}$$

- Variance: The variance measures the average squared deviation of each random variable from their mean.

$$\text{variance: } s_{\text{X}}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- Standard Deviation: The standard deviation is the square root of the variance.

$$\text{standard deviation: } s_{\text{X}} = \sqrt{s_{\text{X}}^2}$$

Dispersion measures

- Interquartile Range (IQR): The interquartile range is the range between the first quartile (25th percentile) and the third quartile (75th percentile) of the random variables.

$$\text{IQR} = Q_3 - Q_1$$

- Mean Absolute Deviation (MAD): The mean absolute deviation calculates the average absolute difference between each random variable and their mean.

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n} |X_i - \bar{X}|$$

Univariate analysis: Asymmetry (skewness)

Definition

Skewness is a measure of the asymmetry of a probability distribution. It quantifies the extent to which the probability distribution of a random variable deviates from being symmetric.

Skewness is typically denoted by the symbol α and is calculated using the third standardized moment. The formula for skewness is as follows:

$$\alpha = \frac{\mathbb{E}[(X - \mathbb{E}(X))^3]}{[\mathbb{E}[(X - \mathbb{E}(X))^2]]^{3/2}}$$

Univariate analysis: Asymetry (skewness)

Skewness interpretability

- $\alpha > 0$ indicates a distribution with a longer tail on the right side, known as right-skewed or positively skewed.
- $\alpha < 0$ indicates a longer tail on the left side, referred to as left-skewed or negatively skewed.
- $\alpha = 0$ indicates a perfectly symmetric distribution.

Univariate analysis: Asymetry (skewness)

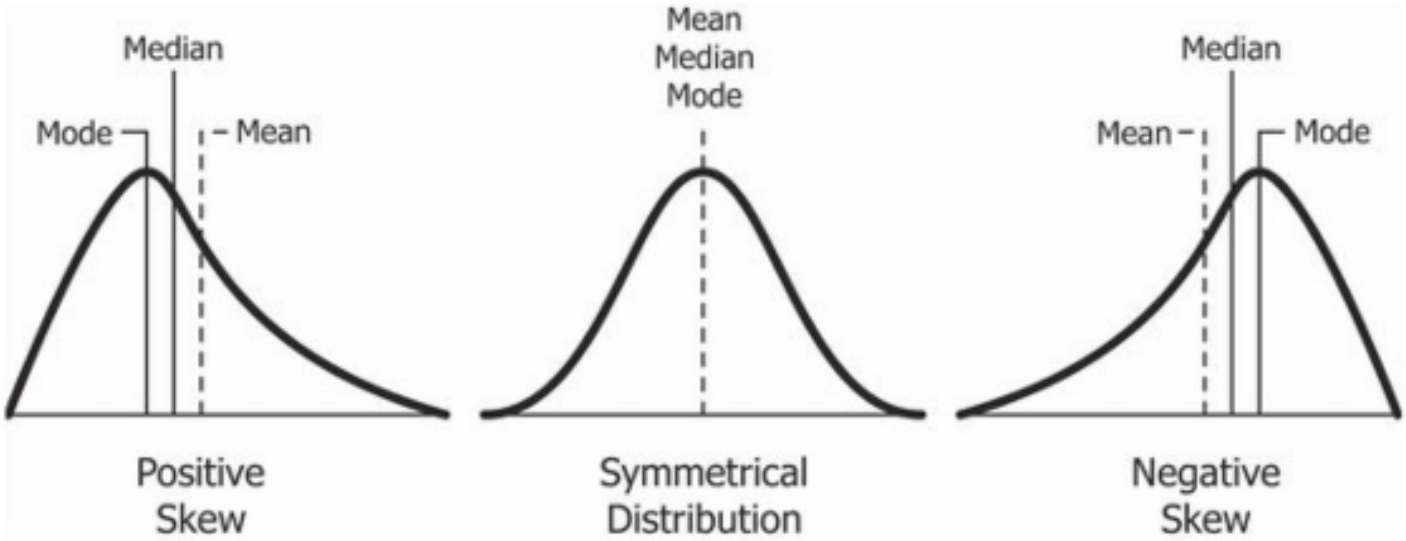

Figure: Skewness example

Univariate analysis: Kurtosis

Definition

Kurtosis is a statistical measure that quantifies the shape of a probability distribution. It captures the degree of peakedness or flatness of a distribution compared to the shape of the normal distribution. Kurtosis uses the fourth moment of a random variable to characterize the distribution's tail behavior.

The formula for kurtosis, denoted by κ, is as follows:

$$\kappa = \frac{\mathbb{E}[(X - \mathbb{E}(X))^4]}{[\mathbb{E}[(X - \mathbb{E}(X))^2]]^2}$$

Univariate analysis: Kurtosis

Kurtosis interpretability

- $\kappa > 0$ indicates heavy tails or distribution with more outliers compared to the normal distribution. This indicates a peaked distribution with potentially thicker tails.
- $\kappa < 0$ indicates light tails or distribution with fewer outliers compared to the normal distribution.
- $\kappa = 0$ indicates a distribution with similar tail behavior as the normal distribution.

Univariate analysis: Kurtosis

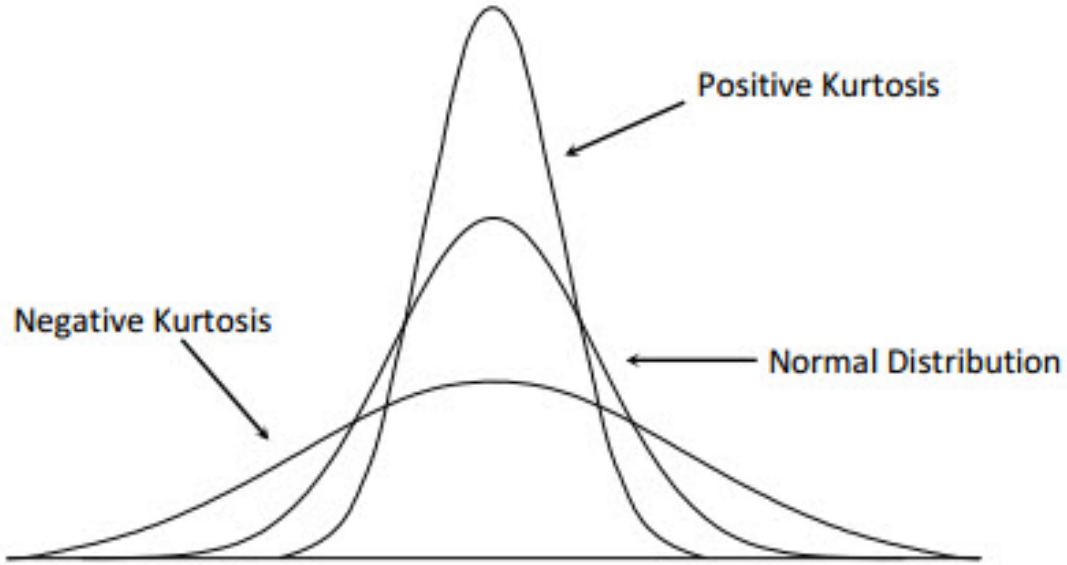

Figure: Kurtosis example

Bivariate analysis

Definition

Bivariate analysis in statistics involves the analysis of two variables simultaneously to understand the relationship between them. It aims to explore patterns, associations, and dependencies between the variables. Several statistical measures and techniques can be used for bivariate analysis.

Bivariate analysis

```python
import pandas as pd
# Load dataset
pd.read_csv("diamonds.csv")
# Show 10 first raws of df
df.head(5)
```

	carat	cut	color	clarity	depth	table	price	x	y	z
0	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75

Figure: Analyze diamonds by their cut, color, clarity, price, and other attributes

Bivariate analysis

```
# Select numerical variables
df1 = df.select_dtypes(include='number')
# Describle numerical variables
df1.describe()
```

	carat	depth	table	price	x	y	z
count	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000
mean	0.797940	61.749405	57.457184	3932.799722	5.731157	5.734526	3.538734
std	0.474011	1.432621	2.234491	3989.439738	1.121761	1.142135	0.705699
min	0.200000	43.000000	43.000000	326.000000	0.000000	0.000000	0.000000
25%	0.400000	61.000000	56.000000	950.000000	4.710000	4.720000	2.910000
50%	0.700000	61.800000	57.000000	2401.000000	5.700000	5.710000	3.530000
75%	1.040000	62.500000	59.000000	5324.250000	6.540000	6.540000	4.040000
max	5.010000	79.000000	95.000000	18823.000000	10.740000	58.900000	31.800000

Figure: Describe numerical variables of diamonds dataset

Bivariate analysis

```
# Select numerical variables
df2 = df.select_dtypes(include='object')
# Describe numerical variables
df2.describe()
```

	cut	color	clarity
count	53940	53940	53940
unique	5	7	8
top	Ideal	G	SI1
freq	21551	11292	13065

Figure: Describe categorical variables of diamonds dataset

Some statistics for bivariate analysis

Let $(X_i, Y_i), i = 1, ..., n$ a set of variable pairs from a given distribution.

- Empirical covariance of observations $((X_1, Y_1), ..., (X_n, Y_n))$, measures how X and Y vary together, indicating the direction and magnitude of the linear relationship.

$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}\bar{Y} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

- Empirical correlation measures the strength and direction of the linear relationship between X and Y.

$$\rho_{XY} = \frac{s_{XY}}{s_X s_Y}$$

Empirical correlation interpretability

- Positive Correlation ($\rho_{XY} > 0$):
  - As one variable increases, the other variable also tends to increase.
  - The strength of the positive correlation depends on the magnitude of the correlation coefficient (ρ_{XY}). A value closer to +1 indicates a stronger positive relationship.
- Negative Correlation ($\rho_{XY} < 0$):
  - As one variable increases, the other variable tends to decrease.
  - The strength of the negative correlation depends on the magnitude of the correlation coefficient (ρ_{XY}). A value closer to -1 indicates a stronger negative relationship.
- No Correlation ($\rho_{XY} \approx 0$):
  - Suggests no linear relationship between the variables, though there might be a nonlinear or non-linear association.
- Perfect Correlation ($\rho_{XY} = \pm 1$):
  - $\rho_{XY} = +1$ indicates a perfect positive linear relationship.
  - $\rho_{XY} = -1$ indicates a perfect negative linear relationship.

Bivariate analysis

```
# Heatmap of df1 (continuous variables)
import seaborn as sns
sns.heatmap(df1.corr())
```

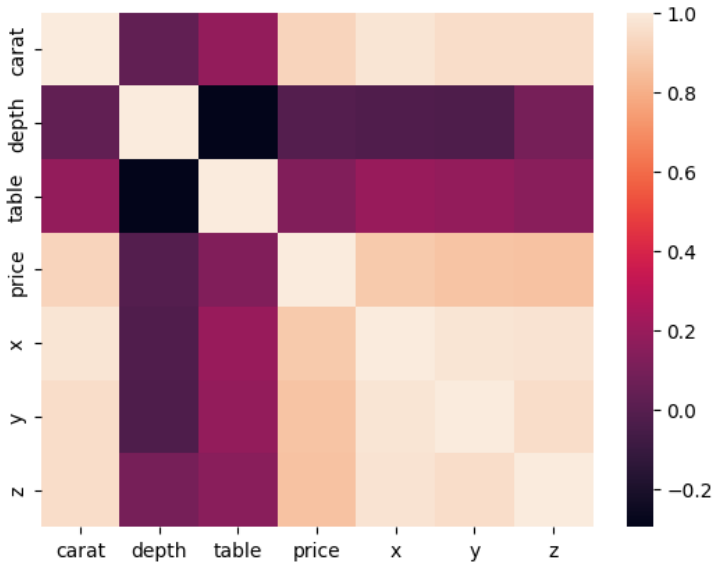

Figure: Correlation of numerical variables in diamonds

Bivariate analysis

```
import seaborn as sns
sns.pairplot(df, x_vars=["depth", "table"],y_vars=["carat", "z"])
```

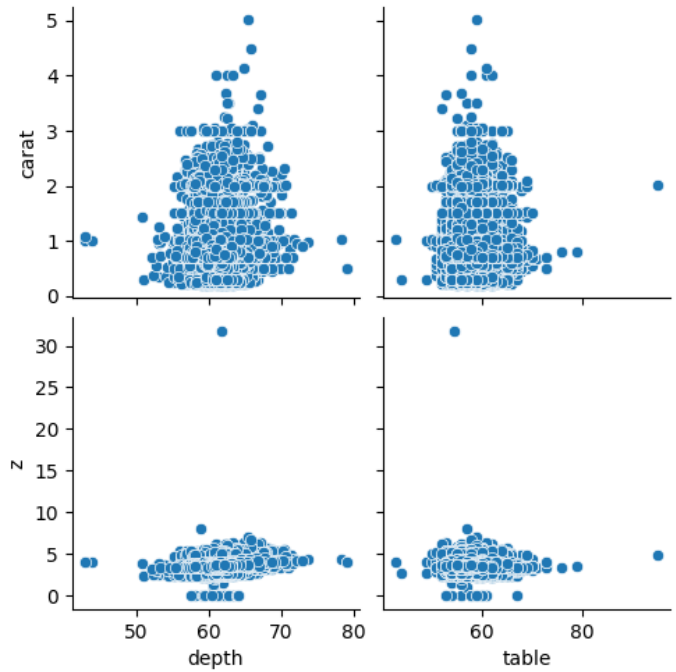

Figure: Scatter of selected numerical variables in diamonds

Bivariate analysis

```
import seaborn as sns
sns.pairplot(df, x_vars=["depth", "table"],y_vars=["carat", "z"], hue="cut")
```

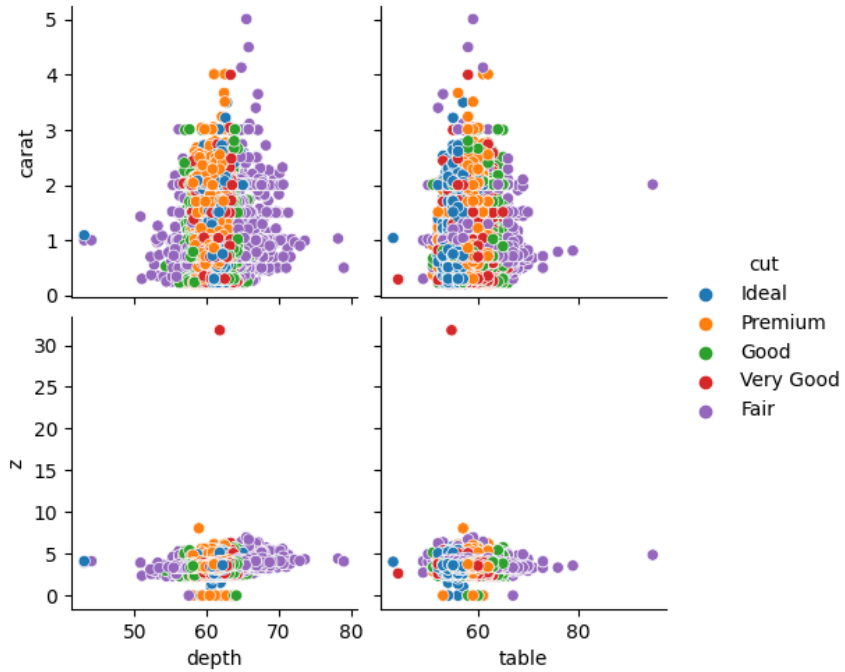

Figure: Scatter of selected numerical variables in diamonds grouping by the categorical variable "cut"

Outline

Machine Learning

Definition

Machine learning is a research field that combines statistics, artificial intelligence, and computer science to extract knowledge from data. It is also known as predictive analytics or statistical learning. In recent years, machine learning has become widespread in everyday life, powering automatic recommendations, personalized online services, and recognition systems. Commercial websites like Facebook, Amazon, and Netflix heavily rely on machine learning algorithms throughout their platforms.

Machine learning: Why Machine learning?

- In the early days, "intelligent" applications relied on handcrafted rules (if-else decisions) for data processing and user interactions, e.g., using word blacklists for spam filters.
- However, this approach had limitations: domain-specific logic and deep human expertise were required, leading to inflexibility in handling new tasks.
- Detecting faces in images highlighted the challenges of handcoding, as computer perception differed from human perception, making rule design difficult.
- In contrast, machine learning enables algorithms to identify face characteristics without explicit rules. For example, providing a program with a large collection of face images allows it to learn what constitutes a face.
- This data-driven approach of machine learning overcomes the manual limitations of rule-based systems and opens up possibilities for tackling complex tasks across various domains.

Machine learning: Problems Machine Learning Can Solve

Machine learning is a powerful field that enables automated decision-making by learning from data. In this context, supervised learning involves providing algorithms with input-output pairs, while unsupervised learning deals with input data only. Let's explore some practical examples of both types of machine learning tasks:

- In supervised learning:
  - Identifying zip codes from handwritten digits on an envelope.
  - Determining whether a tumor is benign based on a medical image.
  - Detecting fraudulent activity in credit card transactions.
- In unsupervised learning:
  - Identifying topics in a set of blog posts.
  - Segmenting customers into groups with similar preferences.
  - Detecting abnormal access patterns to a website.

Machine Learning: Knowing Your Task and Knowing Your Data

Understanding your data and its relevance is essential in machine learning. Avoid randomly applying algorithms; comprehend your dataset before building a model. Tailor algorithms to your problem. Answer questions like:

- Can the data answer my questions?
- How can I frame questions as ML problems?
- Is there enough representative data?
- Are extracted features suitable for predictions?
- How will success be measured?
- How does the ML solution fit into the larger context?

Stay aware of assumptions while building models. Remember, ML is a part of problem-solving, and complex solutions might not solve the right problem.

Machine Learning: Why Python?

Python is a versatile language for data science, with libraries for various tasks like **data loading**, **visualization**, **stats**, and more. It enables quick iteration and interaction through tools like **Jupyter Notebook**. Python's flexibility extends to **GUIs**, **web services**, and integration into existing systems.

Machine Learning: Essential Libraries and Tools

- **scikit-learn**: Machine learning library for various tasks like classification and regression.
- **NumPy**: Core package for numerical computations using arrays and matrices.
- **SciPy**: Extends NumPy with advanced scientific computing features.
- **matplotlib**: Popular plotting library for creating visualizations.
- **pandas**: Data manipulation library for efficient data analysis.
- **mglearn**: Provides tools and datasets for learning and practicing machine learning concepts.

Machine Learning: Essential Libraries and Tools

```
# install packages
pip install numpy scipy matplotlib ipython scikit-learn pandas mglearn
# import some packages
import numpy as np
import sklearn
import scipy as sp
import matplotlib.pyplot as plt
import pandas as pd
import mglearn
```

Machine Learning: A First Application: Classifying Iris Species

The application on the Iris dataset is a classic example in machine learning. The Iris dataset contains measurements of petal and sepal length and width for three species of iris. The goal is to develop a model that can automatically classify iris flowers based on these measurements, showcasing the use of machine learning for classification based on distinct botanical features.

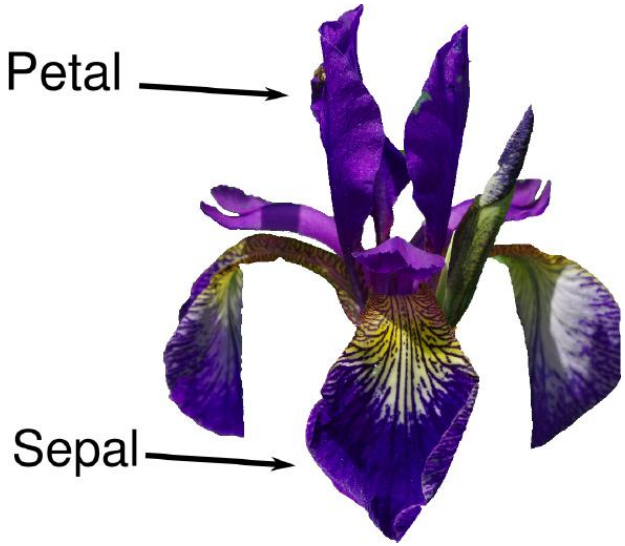

Figure: Parts of the iris flower

Machine Learning: A First Application: Classifying Iris Species

```
# Load the iris dataset
from sklearn.datasets import load_iris
iris_dataset = load_iris()

# Keys on iris dataset
print("Keys of iris_dataset: \n{}".format(iris_dataset.keys()))
```

Machine Learning: A First Application: Classifying Iris Species

Evaluating the model's performance requires separate data to test its generalization. We split our dataset into a training set (75%) and a test set (25%). In scikit-learn, data is represented by capital X and labels by lowercase y. By utilizing the train_test_split function, we can establish these sets and follow standard mathematical conventions for denoting input (X) and output (y) data.

Machine Learning: A First Application: Classifying Iris Species

```
# Split iris dataset to train and test
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(
iris_dataset['data'], iris_dataset['target'], random_state=0)

# Train shape
print("X_train shape: {}".format(X_train.shape))
print("y_train shape: {}".format(y_train.shape))

# Test shape
print("X_test shape: {}".format(X_test.shape))
print("y_test shape: {}".format(y_test.shape))
```

Machine Learning: A First Application: Classifying Iris Species

It is highly recommended to conduct an exploratory analysis of the dataset to gain insights before applying a machine learning algorithm. This approach helps in forming a comprehensive understanding. To illustrate, let's showcase a visualization of the dataset.

```
# create dataframe from data in X_train
# label the columns using the strings in iris_dataset.feature_names
iris_dataframe = pd.DataFrame(X_train, columns=iris_dataset.feature_names)
# create a scatter matrix from the dataframe, color by y_train
grr = pd.plotting.scatter_matrix(iris_dataframe, c=y_train,
figsize=(15, 15), marker='o', hist_kwds={'bins': 20}, s=60,
alpha=.8, cmap=mglearn.cm3)
```

Machine Learning: A First Application: Classifying Iris Species

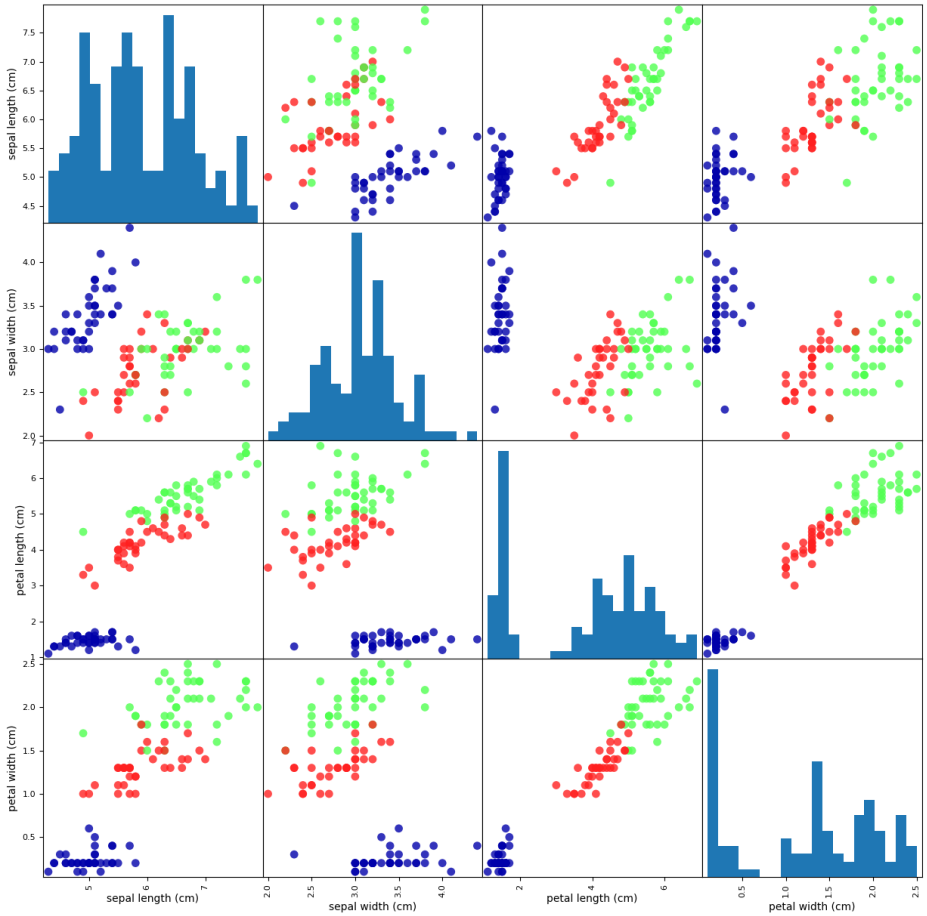

Figure: Pair plot of the Iris dataset, colored by class label

Machine Learning: A First Application: Classifying Iris Species

We will utilize the **k-Nearest Neighbors (KNN) algorithm** to classify the Iris dataset based on its features. KNN is a straightforward yet powerful classification technique. The algorithm's intricacies will be explained in the subsequent slide.

Machine Learning: A First Application: Classifying Iris Species

Algorithm k-Nearest Neighbors

procedure kNN($X, \mathcal{D}, k$)
    $\mathcal{N} \leftarrow$ empty list
    for $\mathbf{x} \in \mathcal{D}$ **do**
        Compute distance $d(\mathbf{x}, X)$
        Add $(\mathbf{x}, d(\mathbf{x}, X))$ to $\mathcal{N}$
    Sort $\mathcal{N}$ by distance
    $\mathcal{N}_k \leftarrow$ first k elements of $\mathcal{N}$
    $\hat{y} \leftarrow$ majority class in $\mathcal{N}_k$
    return $\hat{y}$

Machine Learning: A First Application: Classifying Iris Species

- X: Input instance for prediction
- $\mathcal{D}$: Training dataset with labeled instances
- k: Number of nearest neighbors to consider
- $\mathcal{N}$: List to store neighbors and distances
- $\mathbf{x}$: Instance from training dataset $\mathcal{D}$
- $d(\mathbf{x}, X)$: Distance between $\mathbf{x}$ and X
- $\mathcal{N}_k$: Subset of $\mathcal{N}$ with k nearest neighbors
- $\hat{y}$: Predicted class label for X based on $\mathcal{N}_k$

Machine Learning: A First Application: Classifying Iris Species

Build the k-nearest model:
All machine learning models in scikit-learn are implemented in their own classes, which are called Estimator classes. The k-nearest neighbors classification algorithm is implemented in the KNeighborsClassifier class in the neighbors module.

```
# Import the k-nearest model
from sklearn.neighbors import KNeighborsClassifier
# Create our k-nearest model
knn = KNeighborsClassifier(n_neighbors=1)
```

Machine Learning: A First Application: Classifying Iris Species

Train the k-nearest model on iris dataset :
To build the model on the training set, we call the fit method of the knn object, which takes as arguments the NumPy array X_train containing the training data and the NumPy array y_train of the corresponding training labels:

```
# Build the k-nearest model on the iris training dataset
knn.fit(X_train, y_train)
```

Machine Learning: A First Application: Classifying Iris Species

Making Predictions :
We can now make predictions using this model on new data for which we might not know the correct labels. Imagine we found an iris in the wild with a sepal length of 5 cm, a sepal width of 2.9 cm, a petal length of 1 cm, and a petal width of 0.2 cm.

```
# Put the data into a NumPy array
X_new = np.array([[5, 2.9, 1, 0.2]])
print("X_new.shape: {}".format(X_new.shape))

# Make a prediction
prediction = knn.predict(X_new)
print("Prediction: {}".format(prediction))
print("Predicted target name: {}".format(
iris_dataset['target_names'][prediction]))
```

Machine Learning: A First Application: Classifying Iris Species

Evaluating the Model:
This is where the test set that we created earlier comes in. This data was not used to build the model, but we do know what the correct species is for each iris in the test set.

Therefore, we can make a prediction for each iris in the test data and compare it against its label (the known species). We can measure how well the model works by computing the accuracy, which is the fraction of flowers for which the right species was predicted:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Machine Learning: A First Application: Classifying Iris Species

```
# Make prediction on test dataset
y_pred = knn.predict(X_test)
print("Test set predictions:\n {}".format(y_pred))

# Compute accuracy
print("Test set score: {:.2f}".format(np.mean(y_pred == y_test)))

# Combine prediction and accuracy
print("Test set score: {:.2f}".format(knn.score(X_test, y_test)))
```

Machine Learning: Summary and Outlook

In this chapter, we introduced machine learning applications. We focused on predicting iris species via flower measurements, a supervised task. Using X (features) and y (labels), we split data into train and test sets. We adopted the k-nearest neighbors (KNN) algorithm for classification, achieving 97% accuracy on the test set. This demonstrates the model's ability to predict new data with high confidence.

Machine Learning: Supervised Learning

Definition

Supervised learning is a category of machine learning where the algorithm learns from a labeled dataset, meaning it's provided with input data along with corresponding desired outputs. The goal of supervised learning is to build a model that can accurately map input data to the correct output based on the provided examples.

Machine Learning: Supervised Learning

In supervised machine learning, two primary types of problems stand out: classification and regression.

- **Classification:** Involves categorizing input data into predefined classes. The algorithm learns from labeled data and assigns new data points to the correct categories.

- **Regression:** Focuses on predicting continuous numeric values. Algorithms analyze relationships in training data to make accurate predictions for new data.

Machine Learning: Supervised Learning

In supervised machine learning, achieving the right balance between learning from data and predicting on new data is vital. Concepts like Generalization, Overfitting, and Underfitting play a significant role in this balance.

- **Generalization:**
  - Generalization involves applying learned patterns to new, unseen data.
  - A well-generalized model captures underlying trends without memorizing data.
  - It ensures reliable performance beyond the training set.
- **Overfitting:**
  - Overfitting occurs when a model fits noise and specifics of the training data.
  - Such models perform well on training but poorly on new data.
  - Balancing complexity prevents overfitting and promotes broader applicability.
- **Underfitting:**
  - Underfitting happens when a model is too simplistic to capture data patterns.
  - It leads to poor performance on both training and new data.
  - Addressing underfitting requires increasing model complexity or improving features.

Machine Learning: Supervised Learning

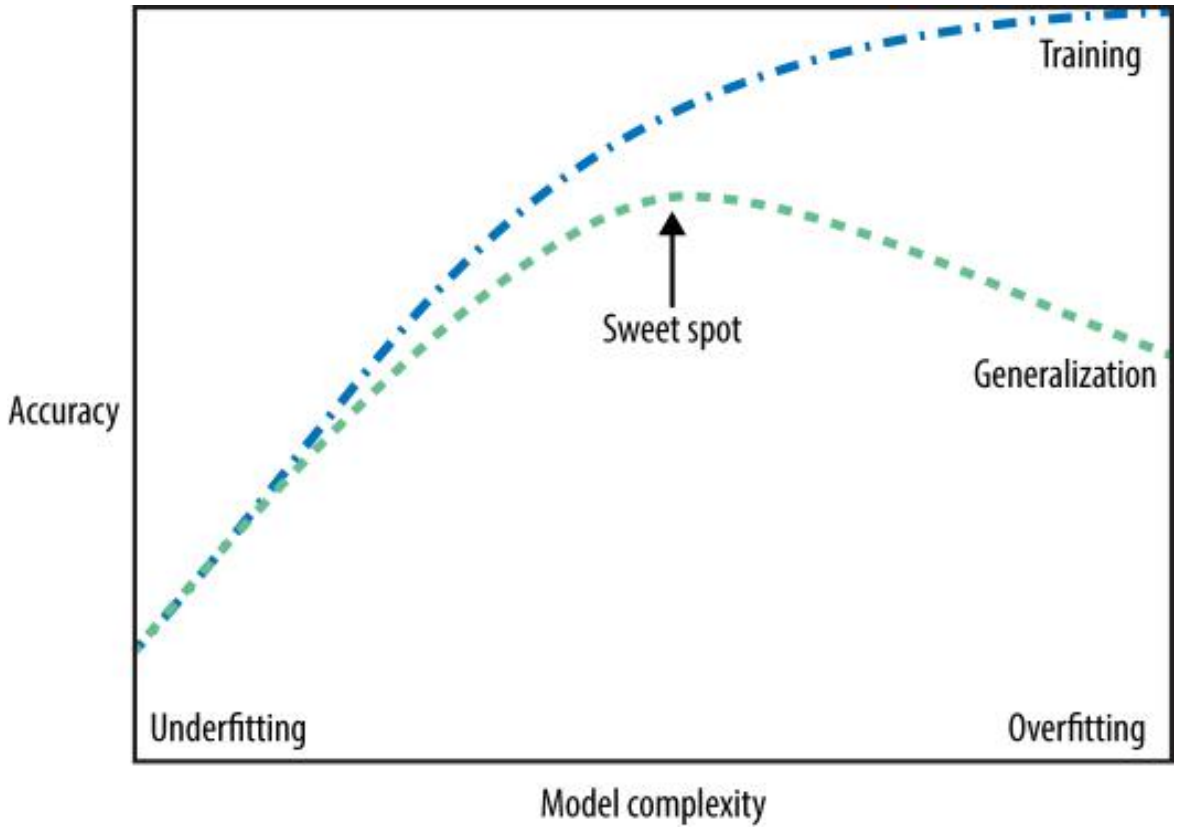

Figure: Trade-off of model complexity against training and test accuracy

Machine Learning: Supervised Learning

Supervised Machine Learning Algorithms:

- k-Nearest Neighbors
  - k-Neighbors classification
  - k-neighbors regression
- Linear Models
  - Linear models for regression
  - Linear models for classification
  - Linear models for multiclass classification
- Naive Bayes Classifiers
- Decision Trees
- Ensembles of Decision Trees
  - Random forests
  - Gradient boosted regression trees (gradient boosting machines)
- Kernelized Support Vector Machines
- Uncertainty Estimates from Classifiers

Supervised Learning: k-Nearest Neighbors

Definition

The k-NN algorithm is arguably the simplest machine learning algorithm. Building the model consists only of storing the training dataset. To make a prediction for a new data point, the algorithm finds the closest data points in the training dataset—its "nearest neighbors".

Supervised Learning: k-Nearest Neighbors: k-Neighbors classification

In its simplest version, the k-NN algorithm only considers exactly one nearest neighbor, which is the closest training data point to the point we want to make a prediction for. The prediction is then simply the known output for this training point.

```
import mglearn
# Plot one-nearest-neighbor model on the forge dataset
mglearn.plots.plot_knn_classification(n_neighbors=1)
```

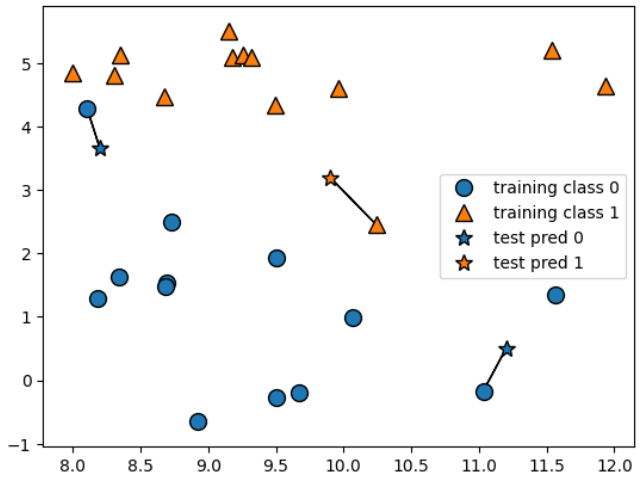

Figure: Predictions made by the one-nearest-neighbor model on the forge dataset

Supervised Learning: k-Nearest Neighbors: k-Neighbors classification

Instead of considering only the closest neighbor, we can also consider an arbitrary number, k, of neighbors. This is where the name of the k-nearest neighbors algorithm comes from. When considering more than one neighbor, we use voting to assign a label.

```
import mglearn
# Plot three-nearest-neighbor model on the forge dataset
mglearn.plots.plot_knn_classification(n_neighbors=3)
```

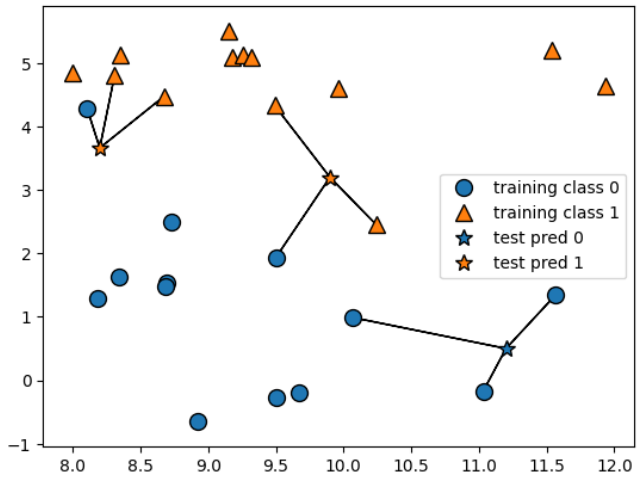

Figure: Predictions made by the three-nearest-neighbors model on the forge dataset

Supervised Learning: k-Nearest Neighbors: k-Neighbors classification

Now let's look at how we can apply the k-nearest neighbors algorithm using scikit-learn. First, we split our data into a training and a test set so we can evaluate generalization performance.

```
import mglearn
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier

X, y = mglearn.datasets.make_forge()
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=0)
clf = KNeighborsClassifier(n_neighbors=3)
clf.fit(X_train, y_train)
# Make prediction on test data
print("Test set predictions: {}".format(clf.predict(X_test)))
# Evaluate how well the model generalizes
print("Test set accuracy: {:.2f}".format(clf.score(X_test, y_test)))
```

Supervised Learning: k-Nearest Neighbors: k-Neighbors classification

Analyzing KNeighborsClassifier:
For two-dimensional datasets, we can also illustrate the prediction for all possible test points in the xy-plane. We color the plane according to the class that would be assigned to a point in this region. This lets us view the decision boundary, which is the divide between where the algorithm assigns class 0 versus where it assigns class 1.

Supervised Learning: k-Nearest Neighbors: k-Neighbors classification

Analyzing KNeighborsClassifier:

```
fig, axes = plt.subplots(1, 3, figsize=(10, 3))
for n_neighbors, ax in zip([1, 3, 9], axes):
        # the fit method returns the object self, so we can instantiate
        # and fit in one line
        clf = KNeighborsClassifier(n_neighbors=n_neighbors).fit(X, y)
        mglearn.plots.plot_2d_separator(clf, X, fill=True, eps=0.5, ax=ax,
        alpha=.4)
        mglearn.discrete_scatter(X[:, 0], X[:, 1], y, ax=ax)
        ax.set_title("{} neighbor(s)".format(n_neighbors))
        ax.set_xlabel("feature 0")
        ax.set_ylabel("feature 1")
        axes[0].legend(loc=3)
```

Supervised Learning: k-Nearest Neighbors: k-Neighbors classification

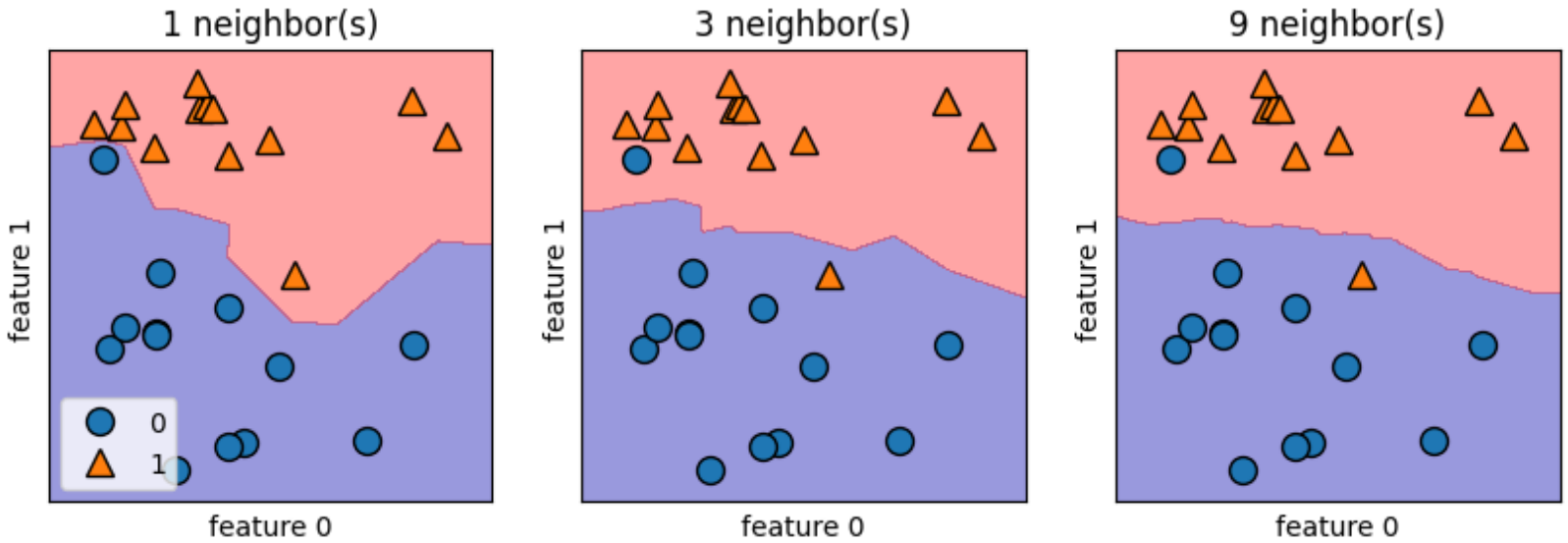

Figure: Decision boundaries created by the nearest neighbors model for different values of n_neighbors

Supervised Learning: k-Nearest Neighbors: k-neighbors regression

There is also a regression variant of the k-nearest neighbors algorithm. Again, let's start by using the single nearest neighbor, this time using the wave dataset. We've added three test data points as green stars on the x-axis. The prediction using a single neighbor is just the target value of the nearest neighbor.

```
import mglearn
mglearn.plots.plot_knn_regression(n_neighbors=1)
```

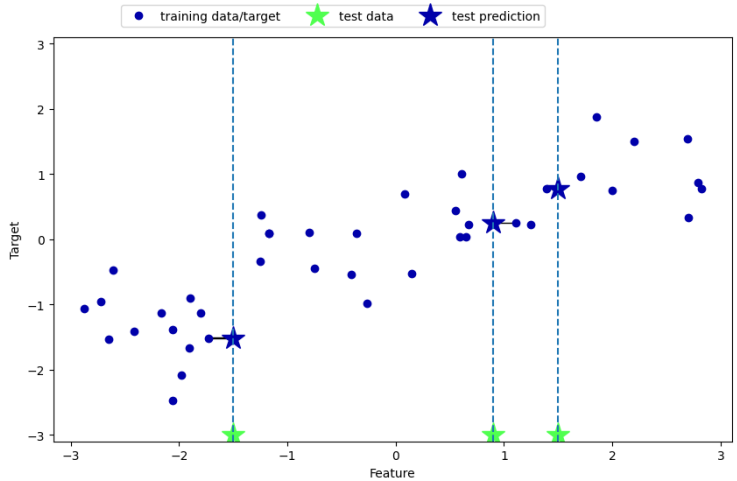

Figure: Predictions made by one-nearest-neighbor regression on the wave dataset

Supervised Learning: k-Nearest Neighbors: k-neighbors regression

Again, we can use more than the single closest neighbor for regression. When using multiple nearest neighbors, the prediction is the average, or mean, of the relevant neighbors.

```
import mglearn
mglearn.plots.plot_knn_regression(n_neighbors=3)
```

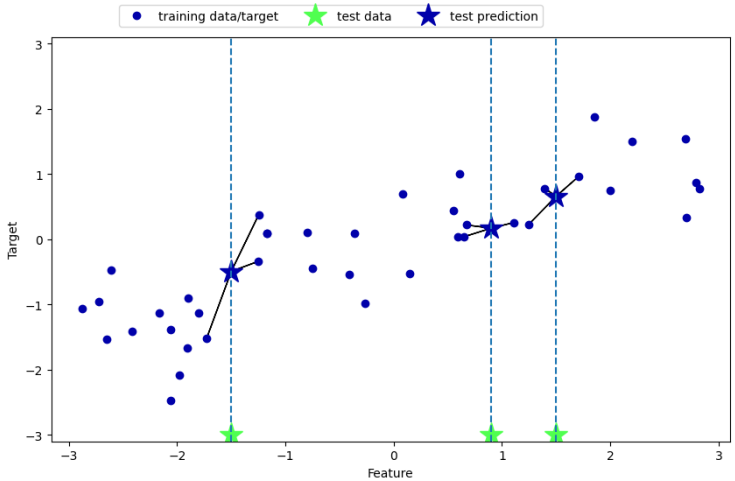

Figure: Predictions made by three-nearest-neighbors regression on the wave dataset

Supervised Learning: k-Nearest Neighbors: k-neighbors regression

The k-nearest neighbors algorithm for regression is implemented in the KNeighbors Regressor class in scikit-learn . It's used similarly to KNeighborsClassifier :

```
import mglearn
from sklearn.neighbors import KNeighborsRegressor
X, y = mglearn.datasets.make_wave(n_samples=40)
# split the wave dataset into a training and a test set
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=0)
# instantiate the model and set the number of neighbors to consider to 3
reg = KNeighborsRegressor(n_neighbors=3)
# fit the model using the training data and training targets
reg.fit(X_train, y_train)
# Make prediction on the test set
print("Test set predictions:\n{}".format(reg.predict(X_test)))
# Evaluate the model using R^2 score
print("Test set R^2: {:.2f}".format(reg.score(X_test, y_test)))
```

Supervised Learning: k-Nearest Neighbors: k-neighbors regression

The coefficient of determination (R-squared) is calculated using the formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Where:
n is the number of data points.
y_i signifies the actual observed value for the i th data point.
$\hat{y}_i$ is the predicted value for the i th data point by the model.
$\bar{y}$ stands for the mean of the observed values.

For our one-dimensional dataset, we can see what the predictions look like for all possible feature values. To do this, we create a test dataset consisting of many points on the line:

Supervised Learning: k-Nearest Neighbors: k-neighbors regression

```
fig, axes = plt.subplots(1, 3, figsize=(15, 4))
line = np.linspace(-3, 3, 1000).reshape(-1, 1)
for n_neighbors, ax in zip([1, 3, 9], axes):
        # make predictions using 1, 3, or 9 neighbors
        reg = KNeighborsRegressor(n_neighbors=n_neighbors)
        reg.fit(X_train, y_train)
        ax.plot(line, reg.predict(line))
        ax.plot(X_train, y_train, '^', c=mglearn.cm2(0), markersize=8)
        ax.plot(X_test, y_test, 'v', c=mglearn.cm2(1), markersize=8)
        ax.set_title(
        "{} neighbor(s)\n train score: {:.2f} test score: {:.2f}".format(
        n_neighbors, reg.score(X_train, y_train),
        reg.score(X_test, y_test)))
        ax.set_xlabel("Feature")
        ax.set_ylabel("Target")
        axes[0].legend(["Model predictions", "Training data/target",
        "Test data/target"], loc="best")
```

Supervised Learning: k-Nearest Neighbors: k-neighbors regression

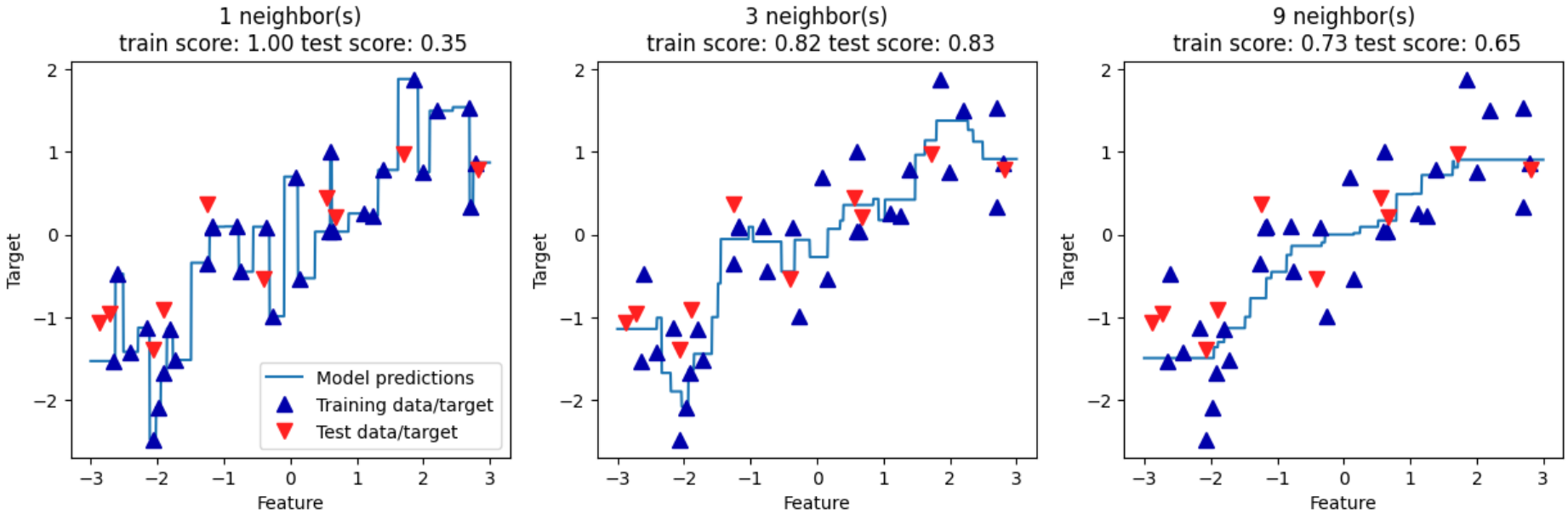

Figure: Comparing predictions made by nearest neighbors regression for different values of n_neighbors

Supervised Learning: k-Nearest Neighbors

Strengths, weaknesses, and parameters:

- **Strengths:**
  - The KNeighbors classifier is easy to understand and implement.
  - It often provides reasonable performance with minimal tuning.
  - It serves as a good baseline model before exploring more complex algorithms.
- **Weaknesses:**
  - Prediction can be slow, especially with large training datasets.
  - It struggles with high-dimensional datasets and sparse features.
  - Not often used in practice due to its limitations and efficiency issues.
- **Parameters:**
  - Important parameters include the number of neighbors and the distance metric.
  - A smaller number of neighbors, like three or five, is often effective.
  - The default Euclidean distance works well for many scenarios.

Supervised Learning: Linear Models

Definition

Linear models are a class of models that are widely used in practice and have been studied extensively in the last few decades, with roots going back over a hundred years. Linear models make a prediction using a linear function of the input features, which we will explain shortly.

For regression, the general prediction formula for a linear model looks as follows:

$$\hat{y} = w[0] * x[0] + w[1] * x[1] + ... + w[p] * x[p] + b$$

Here, $x[0]$ to $x[p]$ denotes the features (in this example, the number of features is p) of a single data point, w and b are parameters of the model that are learned, and $\hat{y}$ is the prediction the model makes.

For a dataset with a single feature, this is:

$$\hat{y} = w[0] * x[0] + b$$

Here, $w[0]$ is the slope and b is the y-axis offset. For more features, w contains the slopes along each feature axis. Alternatively, you can think of the predicted response as being a weighted sum of the input features, with weights (which can be negative) given by the entries of w.

Supervised Learning: Linear Models: Regression

Trying to learn the parameters $w[0]$ and b on our one-dimensional wave dataset might lead to the following line:

```
import mglearn
mglearn.plots.plot_linear_regression_wave()
```

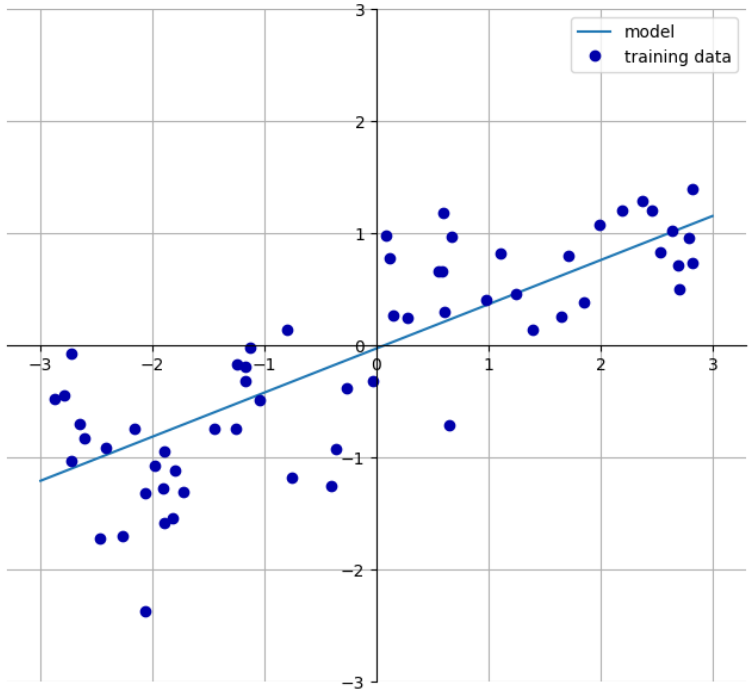

Figure: Predictions of a linear model on the wave dataset: $w[0]$: 0.393906 b : −0.031804

Supervised Learning: Linear Models: Regression

Linear regression, or ordinary least squares (OLS), is the simplest and most classic linear method for regression. Linear regression finds the parameters w and b that minimize the mean squared error between predictions and the true regression targets, y, on the training set. The mean squared error is the sum of the squared differences between the predictions and the true values. Linear regression has no parameters, which is a benefit, but it also has no way to control model complexity.

Supervised Learning: Linear Models: Regression

```
import mglearn
from sklearn.linear_model import LinearRegression
X, y = mglearn.datasets.make_wave(n_samples=60)
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=42)
lr = LinearRegression().fit(X_train, y_train)
print("lr.coef_: {}".format(lr.coef_))
print("lr.intercept_: {}".format(lr.intercept_))
print("Training set score: {:.2f}".format(lr.score(X_train, y_train)))
print("Test set score: {:.2f}".format(lr.score(X_test, y_test)))
```

The "slope" parameters (w), also called weights or coefficients, are stored in the *coef_attribute*, while the offset or intercept (b) is stored in the *intercept_attribute*.

Exercise:
Implement Ridge Regression and Lasso Regression on Boston Housing dataset as alternative methods to enhance the coefficient of determination (R^2) by tuning the regularization parameter (α).

```
import mglearn
from sklearn.linear_model import Ridge
from sklearn.linear_model import Lasso
from sklearn.linear_model import LinearRegression
# Load Boston Housing dataset
X, y = mglearn.datasets.load_extended_boston()
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=0)
# Complete the code
...
```

Supervised Learning: Linear Models: Classification

Linear models are also extensively used for classification. Let's look at binary classification first. In this case, a prediction is made using the following formula:

$$\hat{y} = w[0] * x[0] + w[1] * x[1] + ... + w[p] * x[p] + b > 0$$

The formula looks very similar to the one for linear regression, but instead of just returning the weighted sum of the features, we threshold the predicted value at zero. If the function is smaller than zero, we predict the class –1; if it is larger than zero, we predict the class +1. This prediction rule is common to all linear models for classification. Again, there are many different ways to find the coefficients (w) and the intercept (b).

Supervised Learning: Linear Models: Classification

The two most common linear classification algorithms are logistic regression, and linear support vector machines (linear SVMs). Despite its name, LogisticRegression is a classification algorithm and not a regression algorithm, and it should not be confused with LinearRegression.

Supervised Learning: Linear Models: Classification

We can apply the LogisticRegression and LinearSVC models to the forge dataset, and visualize the decision boundary as found by the linear models:

```
from sklearn.linear_model import LogisticRegression
from sklearn.svm import LinearSVC
X, y = mglearn.datasets.make_forge()
fig, axes = plt.subplots(1, 2, figsize=(10, 3))
for model, ax in zip([LinearSVC(), LogisticRegression()], axes):
    clf = model.fit(X, y)
    mglearn.plots.plot_2d_separator(clf, X, fill=False, eps=0.5,
    ax=ax, alpha=.7)
    mglearn.discrete_scatter(X[:, 0], X[:, 1], y, ax=ax)
    ax.set_title("{}".format(clf.__class__.__name__))
    ax.set_xlabel("Feature 0")
    ax.set_ylabel("Feature 1")
    axes[0].legend()
```

Supervised Learning: Linear Models: Classification

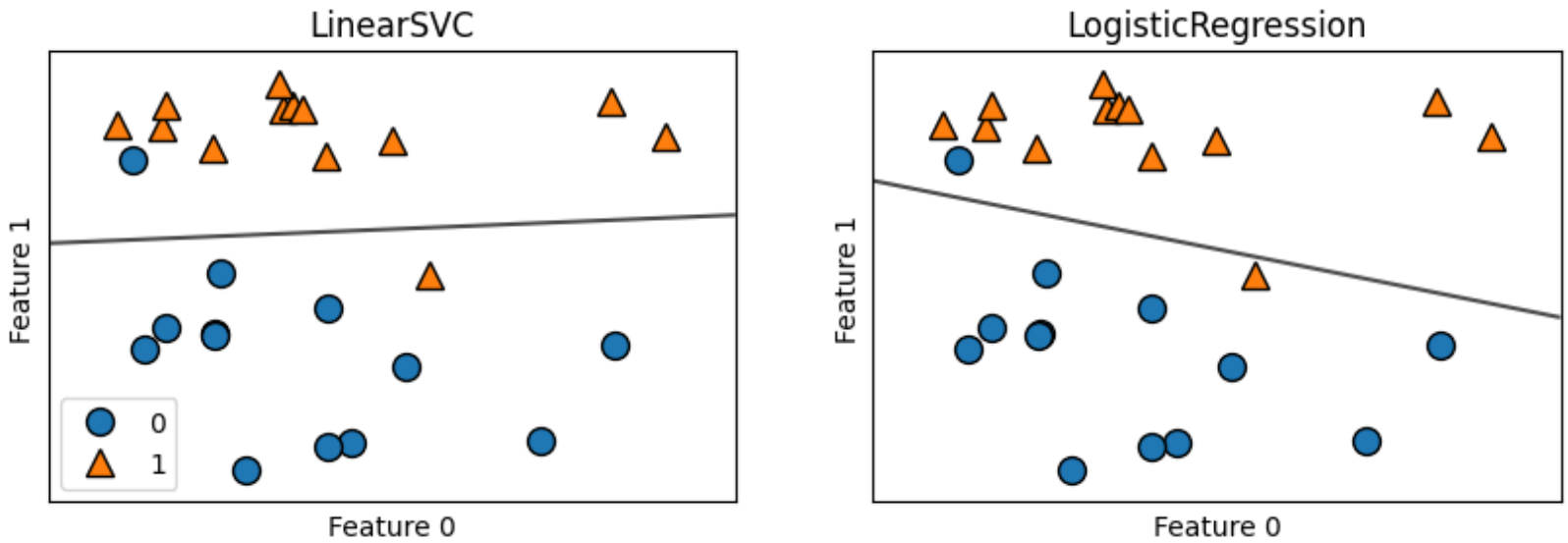

Figure: Decision boundaries of a linear SVM and logistic regression on the forge dataset with the default parameters

The two models come up with similar decision boundaries. Note that both misclassify two of the points. By default, both models apply an L2 regularization, in the same way that Ridge does for regression.

For LogisticRegression and LinearSVC the trade-off parameter that determines the strength of the regularization is called C , and higher values of C correspond to less regularization. In other words, when you use a high value for the parameter C , LogisticRegression and LinearSVC try to fit the training set as best as possible, while with low values of the parameter C , the models put more emphasis on finding a coefficient vector (w) that is close to zero.

The objective of LinearSVC is to find the optimal $\mathbf{w}$ and b that minimize the hinge loss, subject to a regularization term:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\max(0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b))$$

where $\|\mathbf{w}\|^2$ is the L2 norm of the weight vector $\mathbf{w}$, C is the regularization parameter, $\mathbf{x}_i$ is the feature vector of the i-th sample, and y_i is its label.

Supervised Learning: Linear Models: Classification

```
import mglearn
mglearn.plots.plot_linear_svc_regularization()
```

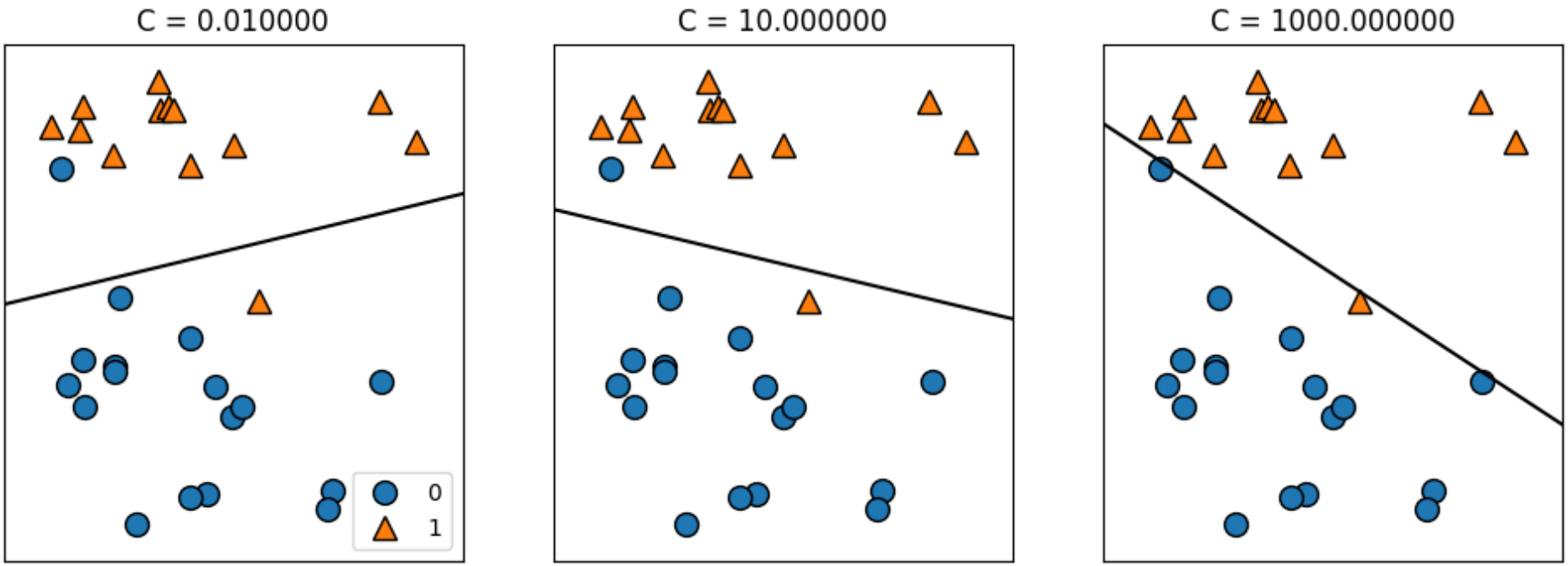

Figure: Decision boundaries of a linear SVM on the forge dataset for different values of C

Exercise:
Explore the impact of different values of the regularization parameter C on Logistic Regression performance in Breast Cancer dataset. Specifically, you will use three different values of: $C = 0.01, C = 1, C = 100$.

```
# Example with default regularization
from sklearn.datasets import load_breast_cancer
cancer = load_breast_cancer()
X_train, X_test, y_train, y_test = train_test_split(
cancer.data, cancer.target, stratify=cancer.target, random_state=42)
logreg = LogisticRegression().fit(X_train, y_train)
print("Training set score: {:.3f}".format(logreg.score(X_train, y_train)))
print("Test set score: {:.3f}".format(logreg.score(X_test, y_test)))
```

Supervised Learning: Linear Models: multiclass classification

Many linear classification models are for binary classification only, and don't extend naturally to the multiclass case (with the exception of logistic regression). A common technique to extend a binary classification algorithm to a multiclass classification algorithm is the one-vs.-rest approach. In the one-vs.-rest approach, a binary model is learned for each class that tries to separate that class from all of the other classes, resulting in as many binary models as there are classes. To make a prediction, all binary classifiers are run on a test point. The classifier that has the highest score on its single class "wins", and this class label is returned as the prediction.

Supervised Learning: Linear Models: multiclass classification

Let's apply the one-vs.-rest method to a simple three-class classification dataset. We use a two-dimensional dataset, where each class is given by data sampled from a Gaussian distribution:

```
import mglearn
from sklearn.datasets import make_blobs
X, y = make_blobs(random_state=42)
mglearn.discrete_scatter(X[:, 0], X[:, 1], y)
plt.xlabel("Feature 0")
plt.ylabel("Feature 1")
plt.legend(["Class 0", "Class 1", "Class 2"])
```

Supervised Learning: Linear Models: multiclass classification

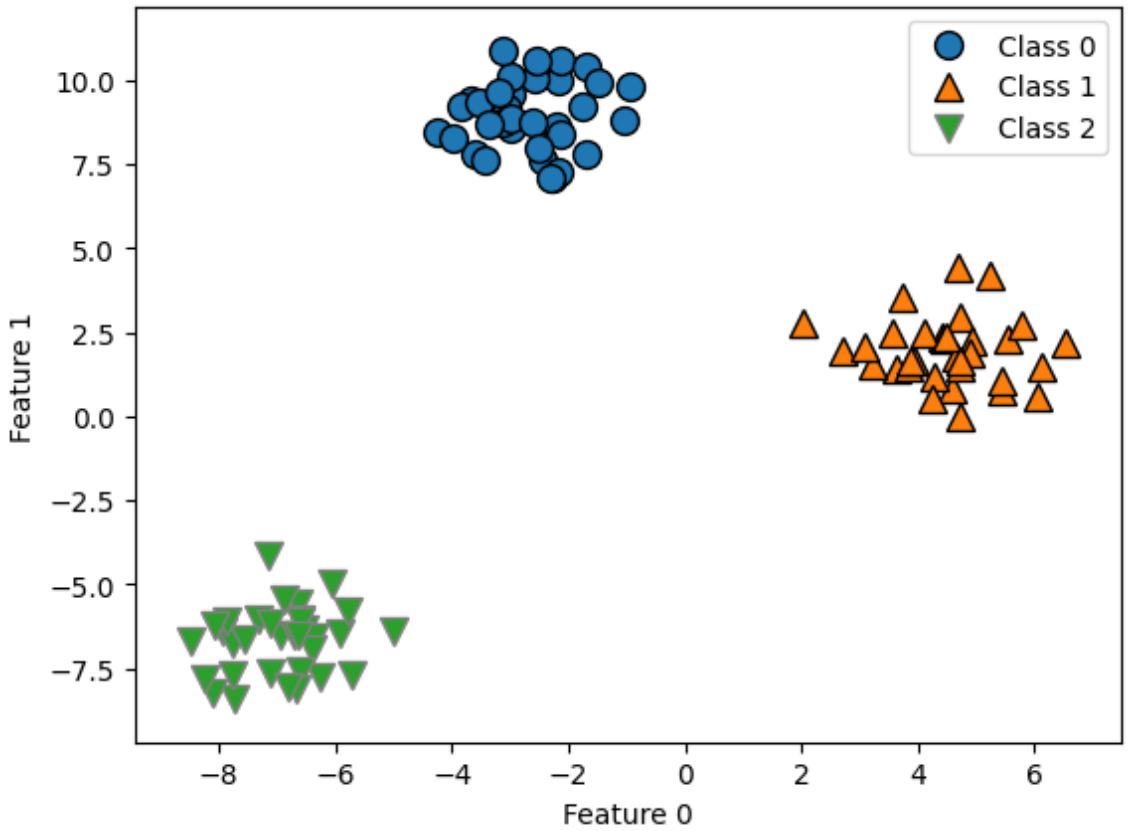

Figure: Two-dimensional toy dataset containing three classes

Supervised Learning: Linear Models: multiclass classification

Now, we train a LinearSVC classifier on the dataset:

```
from sklearn.svm import LinearSVC
import matplotlib.pyplot as plt
import mglearn
linear_svm = LinearSVC().fit(X, y)
print("Coefficient shape: ", linear_svm.coef_.shape)
print("Intercept shape: ", linear_svm.intercept_.shape)

mglearn.discrete_scatter(X[:, 0], X[:, 1], y)
line = np.linspace(-15, 15)
for coef, intercept, color in zip(linear_svm.coef_, linear_svm.intercept_,
['b', 'r', 'g']):
    plt.plot(line, -(line * coef[0] + intercept) / coef[1], c=color)
plt.ylim(-10, 15)
plt.xlim(-10, 8)
plt.xlabel("Feature 0")
plt.ylabel("Feature 1")
plt.legend(['Class 0', 'Class 1', 'Class 2', 'Line class 0', 'Line class 1',
```

Supervised Learning: Linear Models: multiclass classification

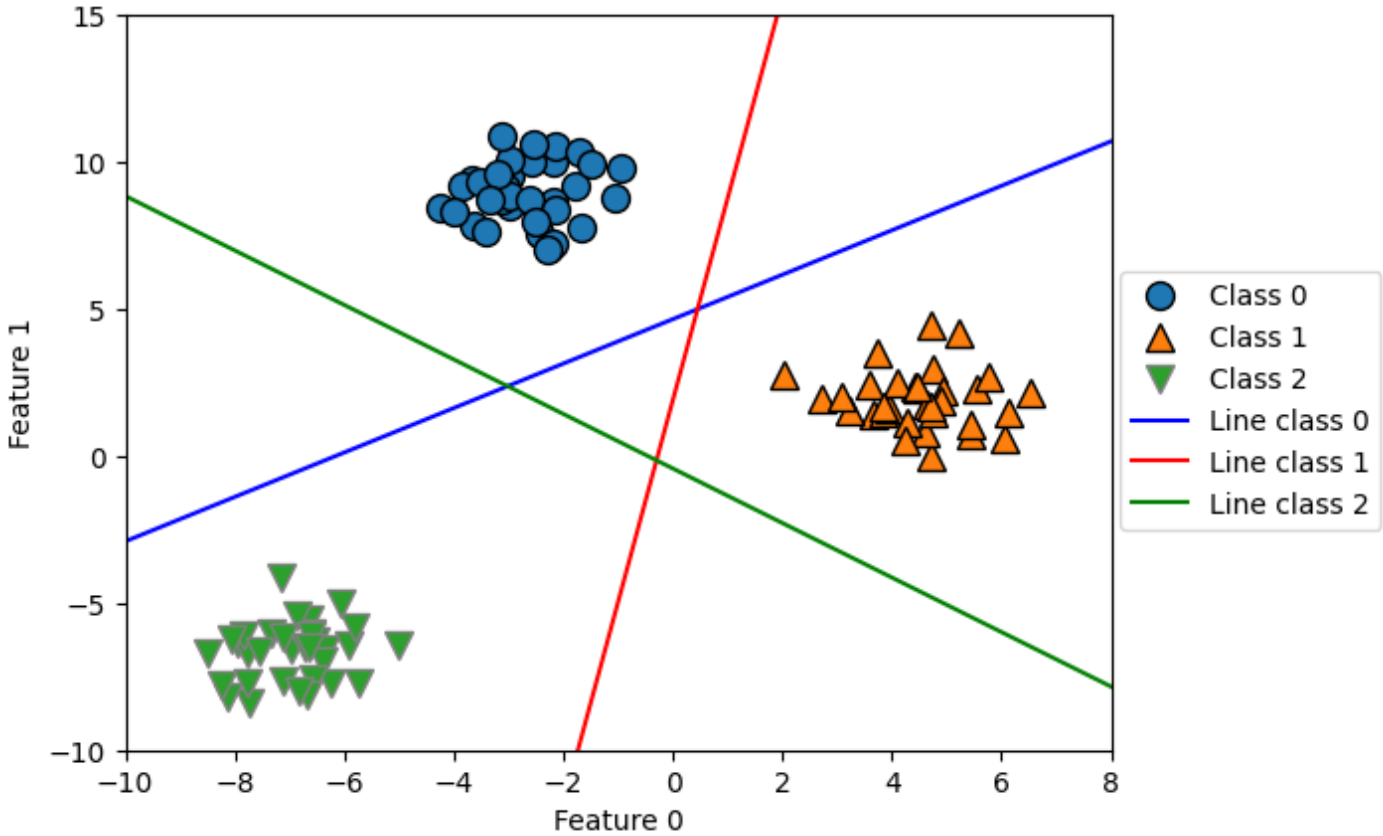

Figure: Decision boundaries learned by the three one-vs.-rest classifiers

Supervised Learning: Linear Models: multiclass classification

The following example shows the predictions for all regions of the 2D space:

```
from sklearn.svm import LinearSVC
import matplotlib.pyplot as plt
import mglearn
mglearn.plots.plot_2d_classification(linear_svm, X, fill=True, alpha=.7)
mglearn.discrete_scatter(X[:, 0], X[:, 1], y)
line = np.linspace(-15, 15)
for coef, intercept, color in zip(linear_svm.coef_, linear_svm.intercept_,
['b', 'r', 'g']):
    plt.plot(line, -(line * coef[0] + intercept) / coef[1], c=color)
plt.legend(['Class 0', 'Class 1', 'Class 2', 'Line class 0', 'Line class 1',
'Line class 2'], loc=(1.01, 0.3))
plt.xlabel("Feature 0")
plt.ylabel("Feature 1")
```

Supervised Learning: Linear Models: multiclass classification

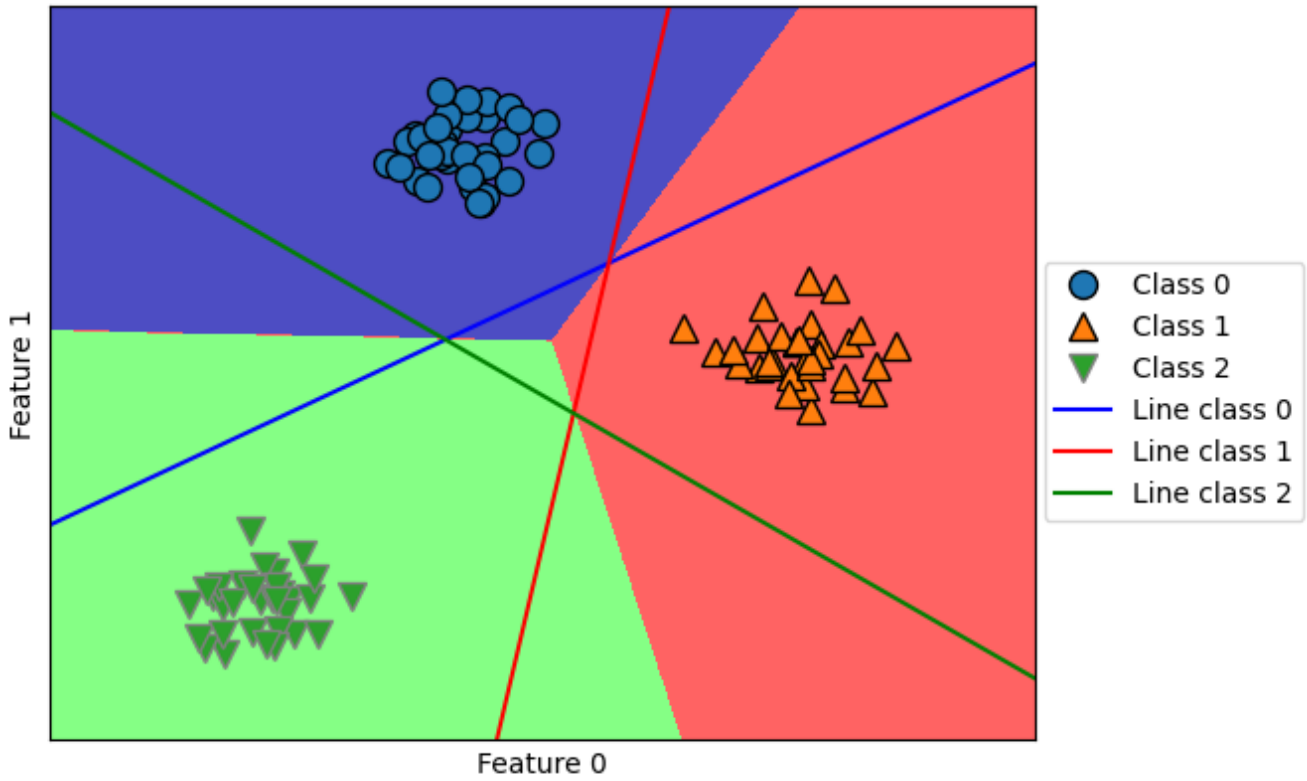

Figure: Multiclass decision boundaries derived from the three one-vs.-rest classifiers

Supervised Learning: Linear Models

Strengths, weaknesses, and parameters:

Strengths:

- Efficient training and prediction.
- Suitable for large datasets and sparse data.
- Scalable options like `solver='sag'`.
- Clear interpretation of predictions.

Weaknesses:

- Interpretation challenges with correlated features.

Parameters:

- Main parameter: Regularization (α in regression, C in LinearSVC and LogisticRegression).
- Larger α or smaller C: Simpler models.
- Tuning α and C important on a logarithmic scale.
- Choose between L1 (sparse important features) and L2 regularization (default).
- L1 aids model interpretability by focusing on key features.

Supervised Learning: Naive Bayes Classifiers

Introduction

Naive Bayes is a probabilistic classification algorithm based on Bayes' Theorem. It assumes that features are conditionally independent given the class label. This simplifying assumption allows for efficient training and prediction. The algorithm is widely used for text classification, spam detection, and more.

Supervised Learning: Naive Bayes Classifiers: Training

Given a dataset with n instances and m features, and C distinct classes $C_1, C_2, \ldots, C_c$, we want to calculate the probability of each class $P(C_i)$ and the conditional probability $P(x_j|C_i)$ for each feature x_j given class C_i.

1. **Calculate Class Priors:**
   - Calculate the total number of instances n and the count of instances n_i belonging to each class C_i.
   - Calculate the prior probability $P(C_i) = \frac{n_i}{n}$
2. **Calculate Conditional Probabilities (with Laplace Smoothing):**
   - For each feature x_j, calculate the number of instances n_{ij} where feature x_j occurs in class C_i.
   - Calculate the conditional probability $P(x_j|C_i) = \frac{n_{ij}+\alpha}{n_i+\alpha \cdot \text{Number of unique feature-values}}$
   - To avoid numerical issues, calculate the logarithm of probabilities: $\log(P(C_i))$ and $\log(P(x_j|C_i))$

Supervised Learning: Naive Bayes Classifiers: Inference

Given a new instance with feature values $x_1, x_2, ..., x_m$, we want to predict its class label C_i.

1. **Calculate Log Posterior Probabilities:**
   - Calculate the log posterior probability for each class C_i using Bayes' Theorem:

$$\log(\mathbb{P}(C_i|x_1, x_2, ..., x_m) = \log(\mathbb{P}(C_i)) + \sum_{j=1}^{m} \mathbb{P}(x_j|C_i)$$

2. **Prediction:**
   - Choose the class C_i that maximizes the log posterior probability:

$$C_{predicted} = argmax_{C_i}(\log(\mathbb{P}(C_i|x_1, x_2, ..., x_m)))$$

Supervised Learning: Naive Bayes Classifiers Implementation

There are three kinds of naive Bayes classifiers implemented in scikit-learn : GaussianNB , BernoulliNB , and MultinomialNB . GaussianNB can be applied to any continuous data, while BernoulliNB assumes binary data and MultinomialNB assumes count data (that is, that each feature represents an integer count of something, like how often a word appears in a sentence). BernoulliNB and MultinomialNB are mostly used in text data classification.

Supervised Learning: Naive Bayes Classifiers

Strengths, weaknesses, and parameters:

- **Strengths:**
  - **Efficiency:** Naive Bayes models are fast to train and predict, making them suitable for large datasets.
  - **Simplicity:** The training procedure is straightforward and easy to understand.
  - **Baseline Model:** Naive Bayes serves as a good baseline model, especially for quick initial assessments.
  - **High-Dimensional Data:** These models perform well on high-dimensional sparse data, such as text data.
- **Weaknesses:**
  - **Independence Assumption:** The "naive" assumption of feature independence may not hold in some cases, leading to potential inaccuracies.
  - **Limited Expressiveness:** Due to the independence assumption, complex relationships between features are not captured.
  - **Accuracy Trade-off:** While often accurate, Naive Bayes might not match the performance of more complex models in certain scenarios.
- **Parameters:**
  - **Alpha:** The smoothing parameter ($0 \leq \alpha \leq 1$) controls model complexity. Higher alpha leads to smoother statistics and simpler models. Tuning alpha can enhance accuracy.

Supervised Learning: Decision Trees

Definition

Decision trees are common models for classification and regression tasks. They use a series of if/else questions to make decisions effectively, similar to a process of elimination. These questions help in distinguishing between different classes or making predictions.

```
import mglearn
mglearn.plots.plot_animal_tree()
```

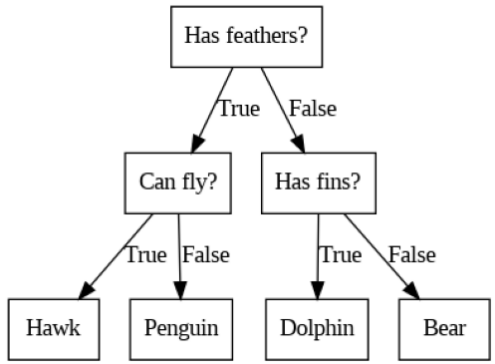

Figure: A decision tree to distinguish among several animals

Supervised Learning: Decision Trees

Building a decision tree involves finding the optimal sequence of if/else questions to efficiently arrive at the correct answer. These questions, known as tests, are used to distinguish between classes or make predictions. In real-world datasets, unlike the animal example, questions often pertain to continuous features such as "Is feature i larger than value a?" We'll walk through this process using the two_moons dataset, consisting of two half-moon shapes with 75 data points in each class.

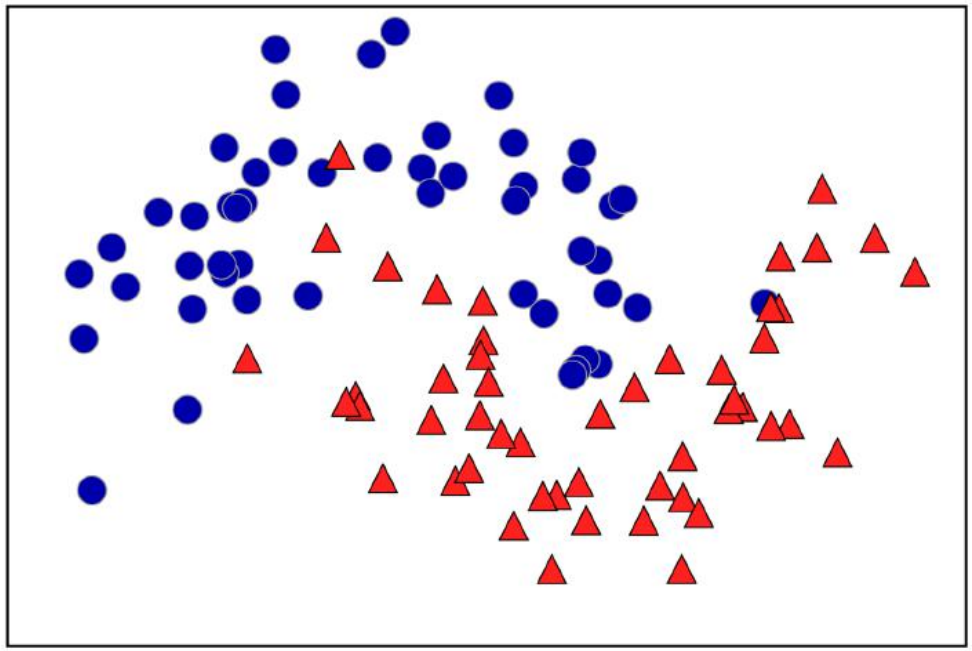

Figure: Two-moons dataset on which the decision tree will be built

Supervised Learning: Decision Trees

The decision tree-building process involves finding the most informative test at each step to separate data points effectively. It begins with an initial test that partitions the dataset based on a specific feature and threshold. The top node represents the entire dataset, and depending on the test outcome, points are assigned to either the left or right child nodes. The goal is to refine these partitions to minimize misclassifications.

The algorithm continues by searching for the most informative tests within each child node, iteratively improving the separation of data points. This process repeats until a stopping criterion is met or until the tree reaches a predefined depth. The result is a hierarchical structure of if/else questions, forming a decision tree that can make predictions based on input features.

Supervised Learning: Decision Trees

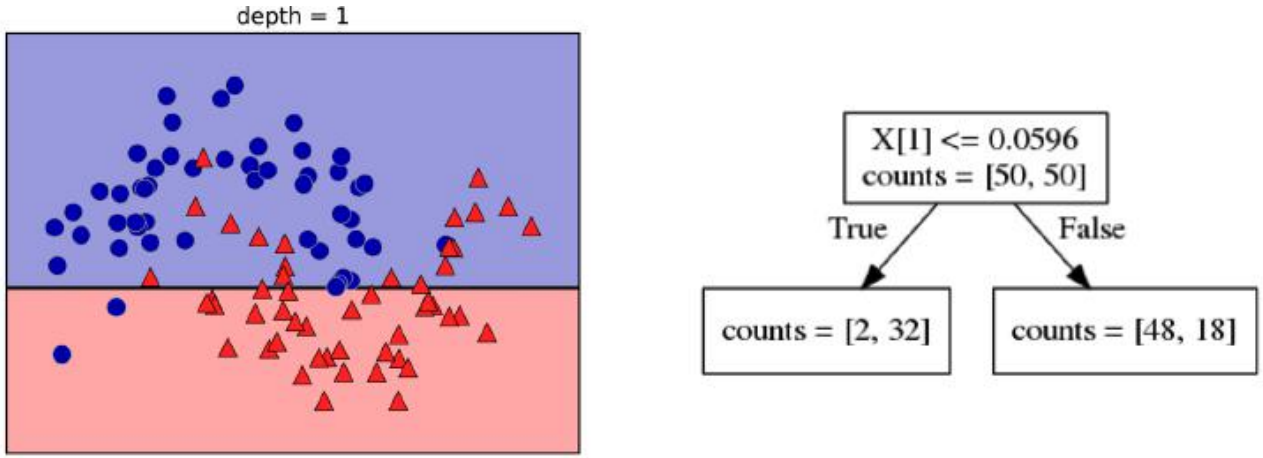

Figure: Decision boundary of tree with depth 1 (left) and corresponding tree (right)

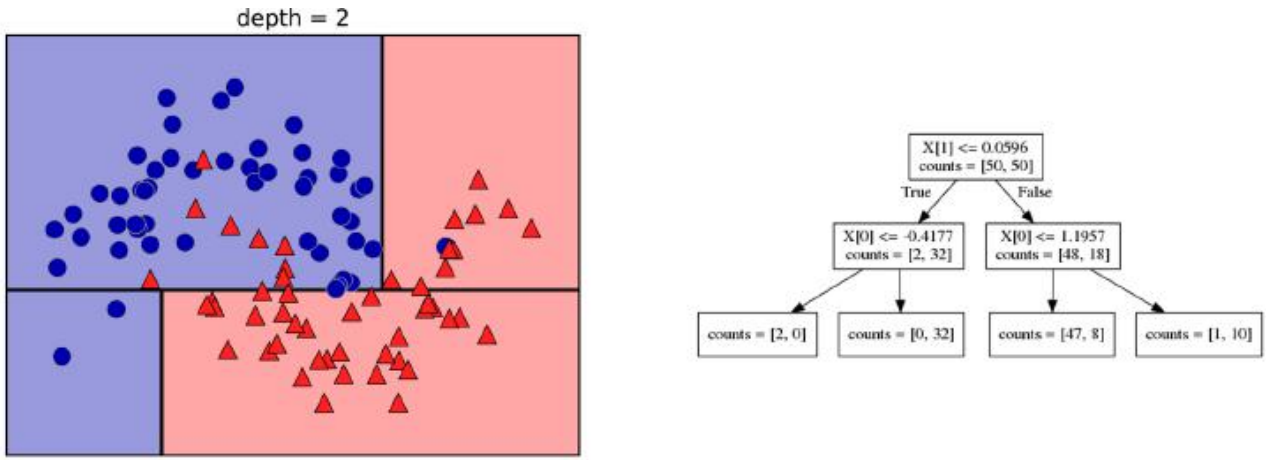

Figure: Decision boundary of tree with depth 2 (left) and corresponding decision tree (right)

Supervised Learning: Decision Trees

The decision tree construction process is recursive, forming a binary tree structure with each node representing a test. These tests split the data along specific axes, resulting in a hierarchical partition. Since each test focuses on a single feature, the partition boundaries are always axis-parallel.
The recursive partitioning continues until each leaf node in the decision tree contains only one unique target value, making it pure. This means that all data points within a pure leaf belong to the same class or have the same regression value.

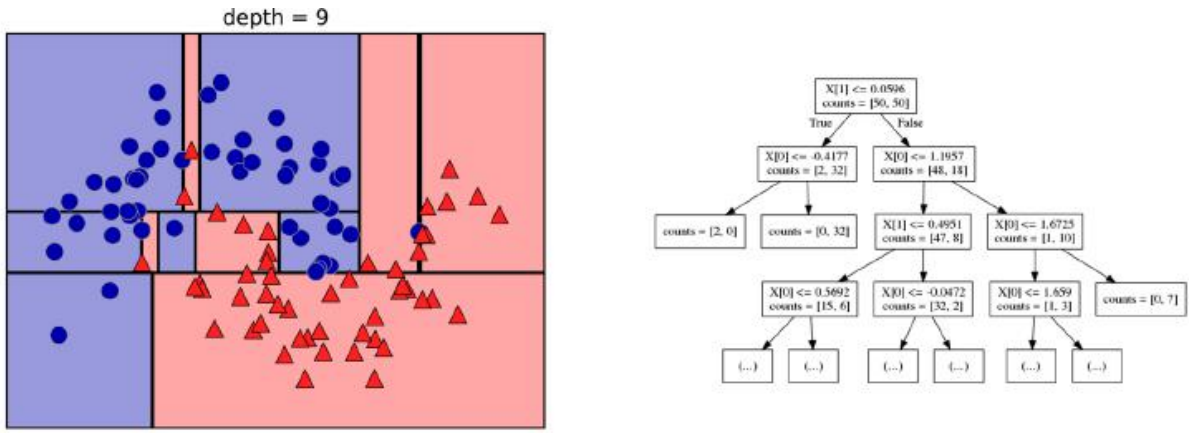

Figure: Decision boundary of tree with depth 9 (left) and part of the corresponding tree (right); the full tree is quite large and hard to visualize

Supervised Learning: Decision Trees

A prediction on a new data point is made by checking which region of the partition of the feature space the point lies in, and then predicting the majority target (or the single target in the case of pure leaves) in that region. The region can be found by traversing the tree from the root and going left or right, depending on whether the test is fulfilled or not.

It is also possible to use trees for regression tasks, using exactly the same technique. To make a prediction, we traverse the tree based on the tests in each node and find the leaf the new data point falls into. The output for this data point is the mean target of the training points in this leaf.

Controlling complexity of decision trees:
Building decision trees by expanding them until all leaves are pure, meaning they perfectly match the training data, often results in overly complex and highly overfit models. Overfitting is evident when decision boundaries incorrectly classify data points and focus too much on outliers, deviating from the expected boundary shapes.

To address overfitting, two common strategies are employed: pre-pruning and post-pruning. Pre-pruning involves setting constraints before constructing the tree, such as limiting the tree's maximum depth, the number of leaves, or the minimum number of data points required to split a node. Post-pruning, on the other hand, builds the tree first and then removes or combines nodes that provide little information. These techniques help create more generalizable decision trees.

Model with pure leaves

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.datasets import load_breast_cancer
from sklearn.model_selection import train_test_split

cancer = load_breast_cancer()
X_train, X_test, y_train, y_test = train_test_split(
cancer.data, cancer.target, stratify=cancer.target, random_state=42)
tree = DecisionTreeClassifier(random_state=0)
tree.fit(X_train, y_train)
print("Accuracy on training set: {:.3f}".format(tree.score(X_train, y_train)))
print("Accuracy on test set: {:.3f}".format(tree.score(X_test, y_test)))
```

Model limiting the depth of the tree decreases overfitting

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.datasets import load_breast_cancer
from sklearn.model_selection import train_test_split

cancer = load_breast_cancer()
X_train, X_test, y_train, y_test = train_test_split(
cancer.data, cancer.target, stratify=cancer.target, random_state=42)
tree = DecisionTreeClassifier(max_depth=4, random_state=0)
tree.fit(X_train, y_train)
print("Accuracy on training set: {:.3f}".format(tree.score(X_train, y_train)))
print("Accuracy on test set: {:.3f}".format(tree.score(X_test, y_test)))
```

Supervised Learning: Decision Trees

We can visualize the tree using the export_graphviz function from the tree module. This writes a file in the .dot file format, which is a text file format for storing graphs. We set an option to color the nodes to reflect the majority class in each node and pass the class and features names so the tree can be properly labeled:

```
from sklearn.tree import export_graphviz
import graphviz

export_graphviz(tree, out_file="tree.dot", class_names=["malignant", "benign"],
feature_names=cancer.feature_names, impurity=True, filled=True)
with open("tree.dot") as f:
    dot_graph = f.read()
graphviz.Source(dot_graph)
```

Supervised Learning: Decision Trees

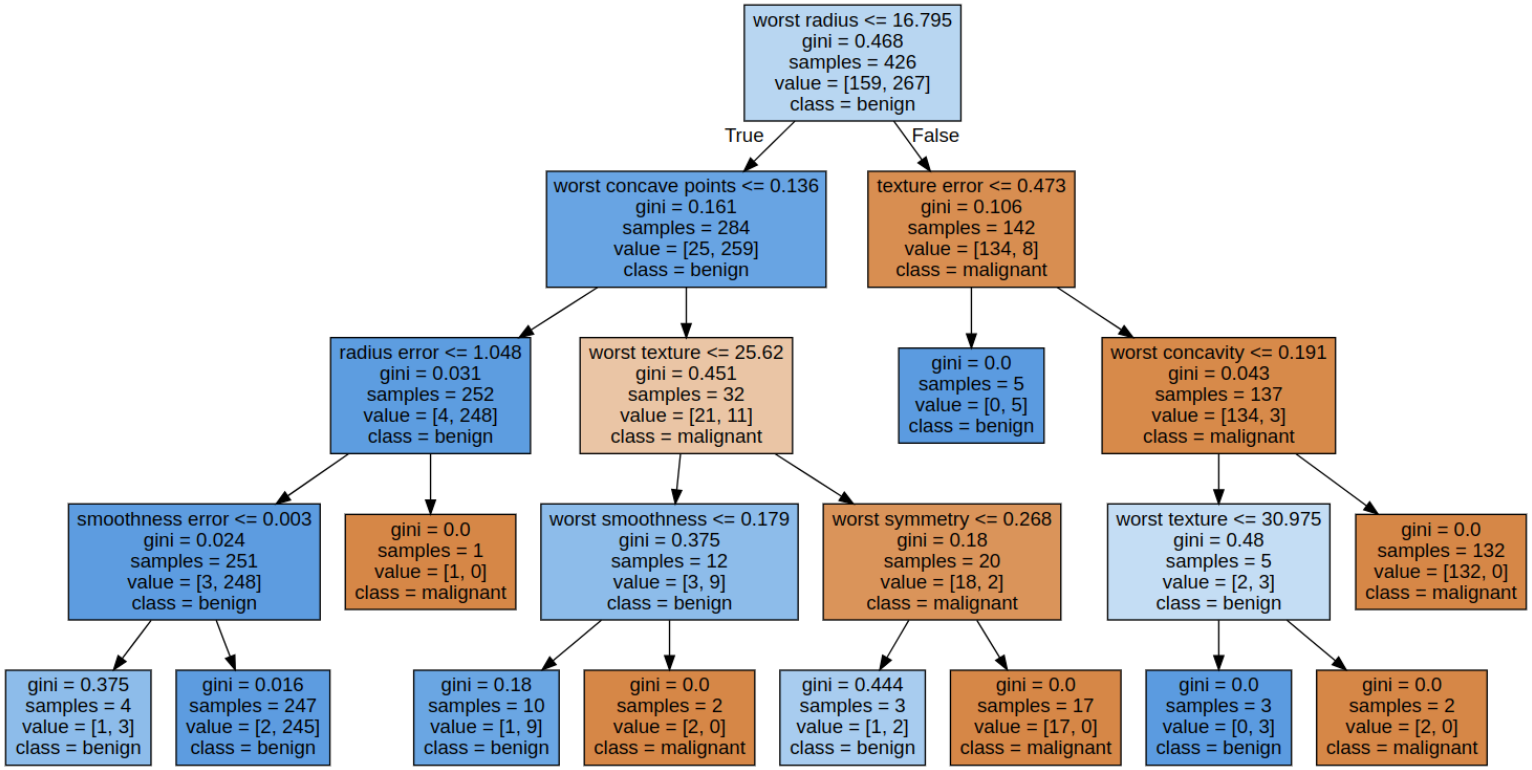

Figure: Visualization of the decision tree built on the Breast Cancer dataset

Feature importance in trees:
Instead of looking at the whole tree, which can be taxing, there are some useful properties that we can derive to summarize the workings of the tree. The most commonly used summary is feature importance, which rates how important each feature is for the decision a tree makes. It is a number between 0 and 1 for each feature, where 0 means "not used at all" and 1 means "perfectly predicts the target." The feature importances always sum to 1:

```
print("Feature importances:\n{}".format(tree.feature_importances_))
```

Supervised Learning: Decision Trees

We can visualize the feature importances in a way that is similar to the way we visualize the coefficients in the linear model:

```
def plot_feature_importances_cancer(model):
    n_features = cancer.data.shape[1]
    plt.barh(range(n_features), model.feature_importances_, align='center')
    plt.yticks(np.arange(n_features), cancer.feature_names)
    plt.xlabel("Feature importance")
    plt.ylabel("Feature")
plot_feature_importances_cancer(tree)
```

Supervised Learning: Decision Trees

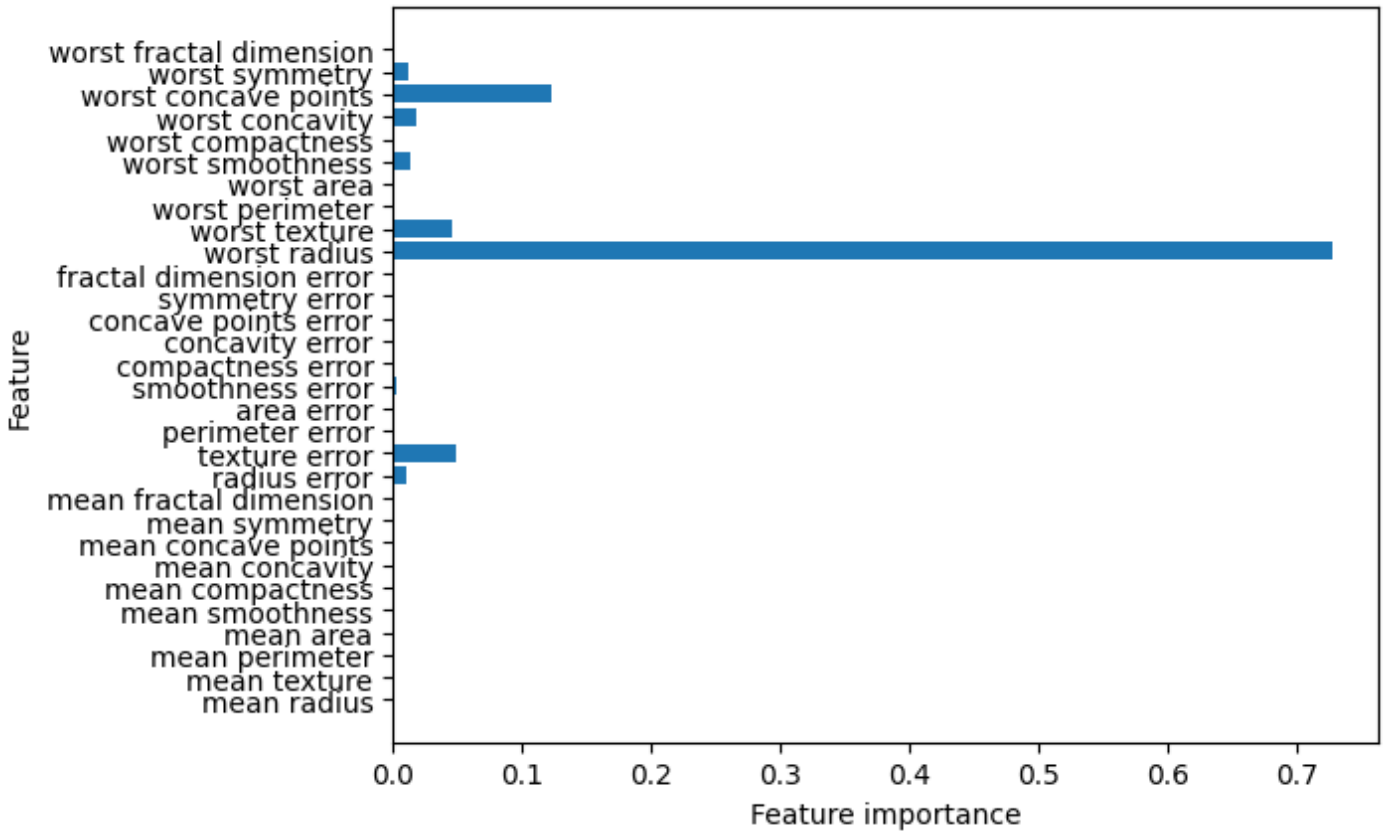

Figure: Feature importances computed from a decision tree learned on the Breast Cancer dataset

Strengths, weaknesses, and parameters:

- **Strengths:**
  - Decision trees are highly interpretable and can be easily visualized and understood, especially for smaller trees.
  - Decision tree algorithms are invariant to data scaling, eliminating the need for preprocessing such as feature normalization or standardization.
  - Effective when dealing with features on different scales or a combination of binary and continuous features.
- **Weaknesses:**
  - Decision trees tend to overfit the training data, leading to poor generalization performance, even with pre-pruning.
- **Parameters:**
  - Pre-pruning parameters can be used to control model complexity:
    - `max_depth`: Limits the maximum depth of the tree.
    - `max_leaf_nodes`: Limits the maximum number of leaf nodes.
    - `min_samples_leaf`: Requires a minimum number of samples in a leaf node to continue splitting.

Supervised Learning: Ensembles of Decision Trees

Definition

Ensembles are methods that combine multiple machine learning models to create more powerful models. There are many models in the machine learning literature that belong to this category, but there are two ensemble models that have proven to be effective on a wide range of datasets for classification and regression, both of which use decision trees as their building blocks: random forests and gradient boos- ted decision trees.

Supervised Learning: Ensembles of Decision Trees: Random Forest

Definition

Random forests are an ensemble learning method that combines multiple decision trees to improve model accuracy and reduce overfitting. They utilize bootstrapping (sample selection) and random feature selection to create diverse trees, and the final prediction in classification tasks is based on a majority vote from individual trees. In regression, predictions are averaged. Random forests are known for their robustness, parallelizability, and feature importance estimation, making them a popular choice for various machine learning tasks.

Supervised Learning: Ensembles of Decision Trees: Random Forest

Building random forests:
To illustrate the process, consider a dataset with samples ['a', 'b', 'c', 'd']. A bootstrap sample may look like ['b', 'd', 'd', 'c'], and another could be ['d', 'a', 'd', 'a']. Decision trees in the random forest are built using a slightly modified algorithm. Instead of considering all features at each node, the algorithm randomly selects a subset of features, controlled by max_features, for making splits. This randomness ensures that each tree in the forest is different.

A critical parameter is max_features. High max_features makes trees more similar, fitting the data using distinctive features, while low max_features makes trees diverse, potentially requiring greater depth to fit the data.

For predictions, regression results are averaged across trees. In classification, a "soft voting" strategy is used, where each tree provides class probabilities. These probabilities are averaged, and the class with the highest probability is the final prediction. This ensemble technique reduces overfitting and enhances model robustness and predictive power.

Supervised Learning: Ensembles of Decision Trees: Random Forest

Analyzing random forests:

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.datasets import make_moons
X, y = make_moons(n_samples=100, noise=0.25, random_state=3)
X_train, X_test, y_train, y_test = train_test_split(X, y, stratify=y,
random_state=42)
forest = RandomForestClassifier(n_estimators=5, random_state=2)
forest.fit(X_train, y_train)
# Visualization
fig, axes = plt.subplots(2, 3, figsize=(20, 10))
for i, (ax, tree) in enumerate(zip(axes.ravel(), forest.estimators_)):
    ax.set_title("Tree {}".format(i))
    mglearn.plots.plot_tree_partition(X_train, y_train, tree, ax=ax)
mglearn.plots.plot_2d_separator(forest, X_train, fill=True, ax=axes[-1, -1],
alpha=.4)
axes[-1, -1].set_title("Random Forest")
mglearn.discrete_scatter(X_train[:, 0], X_train[:, 1], y_train)
```

Supervised Learning: Ensembles of Decision Trees: Random Forest

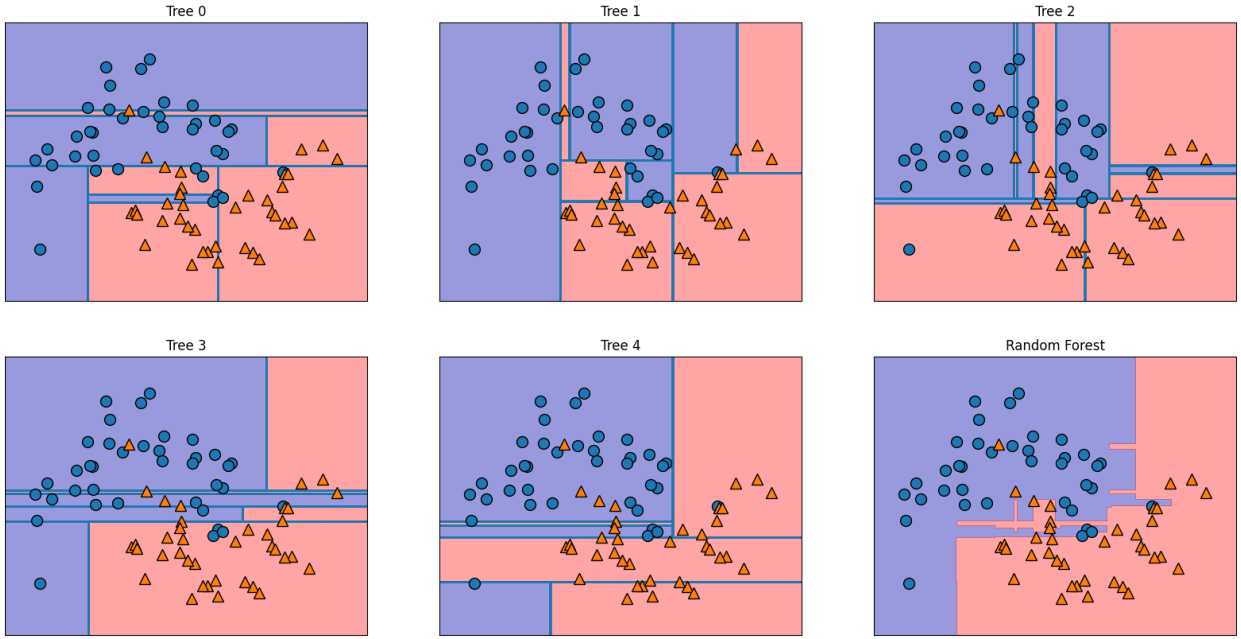

Figure: Decision boundaries found by five randomized decision trees and the decision boundary obtained by averaging their predicted probabilities

Supervised Learning: Ensembles of Decision Trees: Random Forest

Analyzing random forests: As another example, let's apply a random forest consisting of 100 trees on the Breast Cancer dataset:

```
X_train, X_test, y_train, y_test = train_test_split(
cancer.data, cancer.target, random_state=0)
forest = RandomForestClassifier(n_estimators=100, random_state=0)
forest.fit(X_train, y_train)
print("Accuracy on training set: {:.3f}".format(forest.score(X_train, y_train)))
print("Accuracy on test set: {:.3f}".format(forest.score(X_test, y_test)))
```

The random forest gives us an accuracy of 97%, better than the linear models or a single decision tree, without tuning any parameters. We could adjust the max_features setting, or apply pre-pruning as we did for the single decision tree. However, often the default parameters of the random forest already work quite well.

Supervised Learning: Ensembles of Decision Trees: Random Forest

Analyzing random forests: Similarly to the decision tree, the random forest provides feature importances, which are computed by aggregating the feature importances over the trees in the forest. Typically, the feature importances provided by the random forest are more reliable than the ones provided by a single tree.

```
plot_feature_importances_cancer(forest)
```

Supervised Learning: Ensembles of Decision Trees: Random Forest

Analyzing random forests:

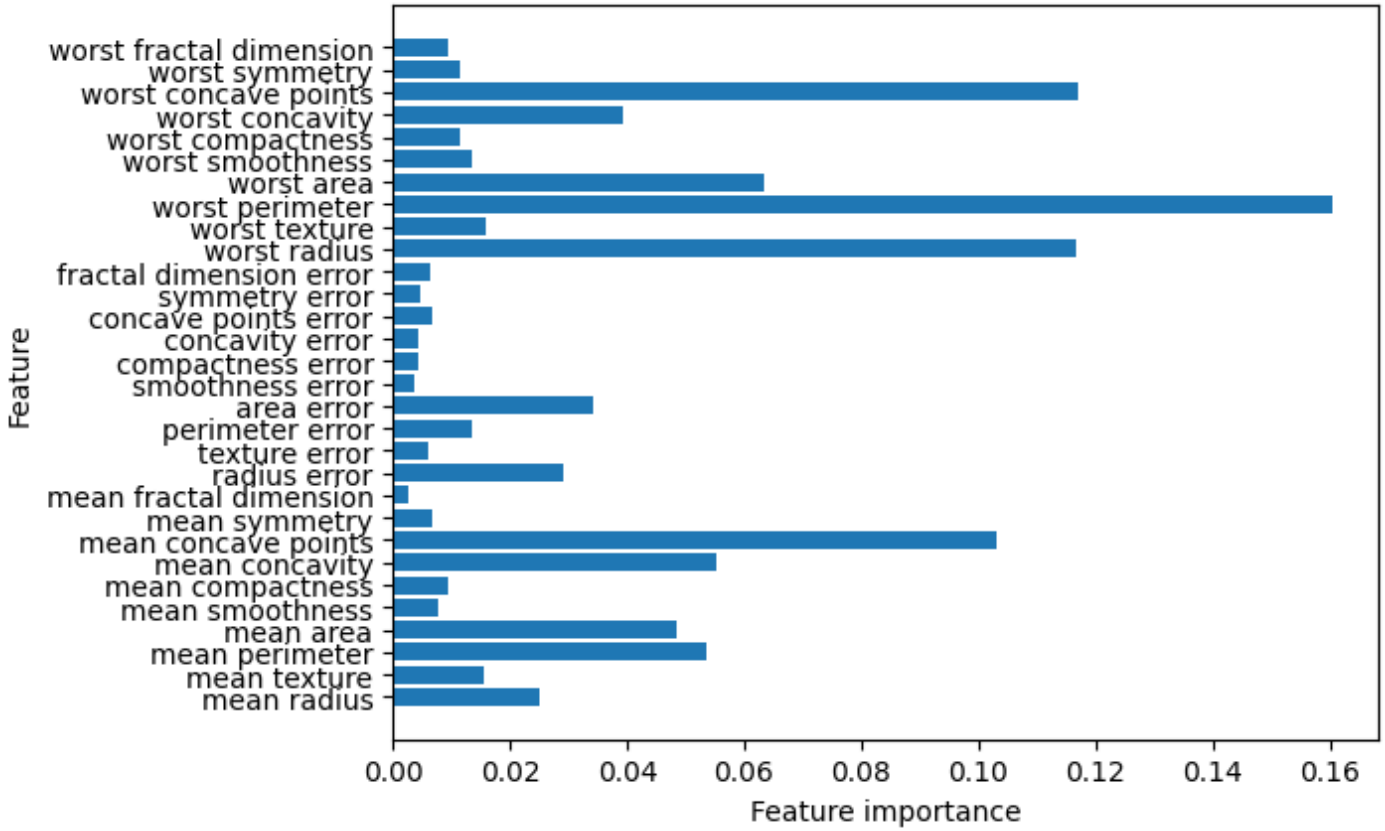

Figure: Feature importances computed from a random forest that was fit to the Breast Cancer dataset

Supervised Learning: Ensembles of Decision Trees: Random Forest

Strengths, Weaknesses, and parameters:

- **Strengths:**
  - They often work effectively without extensive parameter tuning.
  - No need for data scaling, as they are invariant to feature scaling.
  - Parallelizable on multi-core processors, making them suitable for large datasets.
  - Easily interpretable decision-making process for a single decision tree.
- **Weaknesses:**
  - Not suitable when a compact decision-making representation is needed, as they consist of many trees.
  - Interpretability can be challenging with numerous deep trees.
  - Sensitivity to random state settings; different states can result in varying models.
  - Less suitable for high-dimensional sparse data, where linear models might be more appropriate.
  - Requires more memory and is slower compared to linear models.

Supervised Learning: Ensembles of Decision Trees: Random Forest

Strengths, Weaknesses, and parameters:

- **Parameters:**
  - `n_estimators`: Larger values are generally better for better ensemble robustness but require more time and memory.
  - `max_features`: Controls randomness in each tree. Default values ($\sqrt{\text{n\_features}}$ for classification, $\log_2(\text{n\_features})$ for regression) often work well.
  - Pre-pruning options like `max_depth` can be used to limit tree depth.
  - For reproducible results, setting `random_state` is crucial.
  - Consider `max_leaf_nodes` for improved performance and reduced resource requirements.

Supervised Learning: Ensembles of Decision Trees: Gradient boosted regression trees (gradient boosting machines)

Definition

Gradient Boosted Regression Trees, suitable for regression and classification, sequentially build trees to correct errors from previous ones. They are memory-efficient and faster for predictions. By combining many simple models, they achieve high accuracy. Adjust key parameters like learning rate and the number of trees (n_estimators) for optimal results.

In gradient boosting, the learning_rate parameter influences how aggressively each tree corrects the errors of its predecessors. A higher learning rate allows for more substantial corrections and leads to more complex models.

Supervised Learning: Ensembles of Decision Trees: Gradient boosted regression trees

Require: Training dataset $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, M iterations, learning rate ν, max tree depth d, loss function $L(F, y)$ (MSE or Cross-Entropy) test data $\{x_1^*, x_2^*, \ldots, x_k^*\}$

Ensure: Predicted target values $\{y_1^*, y_2^*, \ldots, y_k^*\}$

1: **Initialization:**

2: Initialize ensemble model: $F_0(x) = \text{argmin}_\gamma \sum_{i=1}^{n} L(\gamma, y_i)$, γ the initial prediction

3: **for** $m = 1$ **to** M **do**

4: **Step 1: Compute Negative Gradient:**

5: Calculate negative gradient r_{im} w.r.t. $F_{m-1}(x_i)$ for each training example x_i:

$$r_{im} = -\left[\frac{\partial L(F_{m-1}(x_i), y_i)}{\partial F_{m-1}(x_i)}\right]_{F_{m-1}(x_i)=F_{m-1}(x_i)}$$

6: **Step 2: Fit a Regression Tree:**

7: Train regression tree $h_m(x)$ using r_{im} as target variable and input features x_i as predictors. Max tree depth is d.

Supervised Learning: Ensembles of Decision Trees: Gradient boosted regression trees

1: **for** $m = 1$ **to** M **do**
2: ...
3: **Step 3: Update Ensemble Model:**
4: Update ensemble model $F_m(x)$ by adding prediction of newly trained tree scaled by ν:

$$F_m(x) = F_{m-1}(x) + \nu \cdot h_m(x)$$

5: **Step 4: Update Loss Function:**
6: Update loss function $L(F, y)$ by adding loss contributed by newly added tree:

$$L(F, y) = L(F_{m-1} + \nu \cdot h_m, y)$$

7: **Termination:**
8: Repeat Steps 1 to 4 for M iterations or until predefined stopping criterion is met during training.

Supervised Learning: Ensembles of Decision Trees: Gradient boosted regression trees

1: **Inference:**
2: **for** each test sample x_i^* **do**
3: **Step 5: Predict Target Value:**
4: Predict target value y_i^* for test sample x_i^* using final ensemble model $F_M(x_i^*)$:

$$y_i^* = F_M(x_i^*)$$

5: **Output:**
6: Predicted target values $\{y_1^*, y_2^*, \ldots, y_k^*\}$

Supervised Learning: Ensembles of Decision Trees: Gradient boosted regression trees

Example of using GradientBoostingClassifier on the Breast Cancer dataset. By default, 100 trees of maximum depth 3 and a learning rate of 0.1 are used:

```
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.model_selection import train_test_split
from sklearn.datasets import load_breast_cancer

cancer = load_breast_cancer()
X_train, X_test, y_train, y_test = train_test_split(
cancer.data, cancer.target, random_state=0)
gbrt = GradientBoostingClassifier(random_state=0)
gbrt.fit(X_train, y_train)
print("Accuracy on training set: {:.3f}".format(gbrt.score(X_train, y_train)))
print("Accuracy on test set: {:.3f}".format(gbrt.score(X_test, y_test)))
```

Supervised Learning: Ensembles of Decision Trees: Gradient boosted regression trees

Visualize the feature importances

```
gbrt = GradientBoostingClassifier(random_state=0, max_depth=1)
gbrt.fit(X_train, y_train)
plot_feature_importances_cancer(gbrt)
```

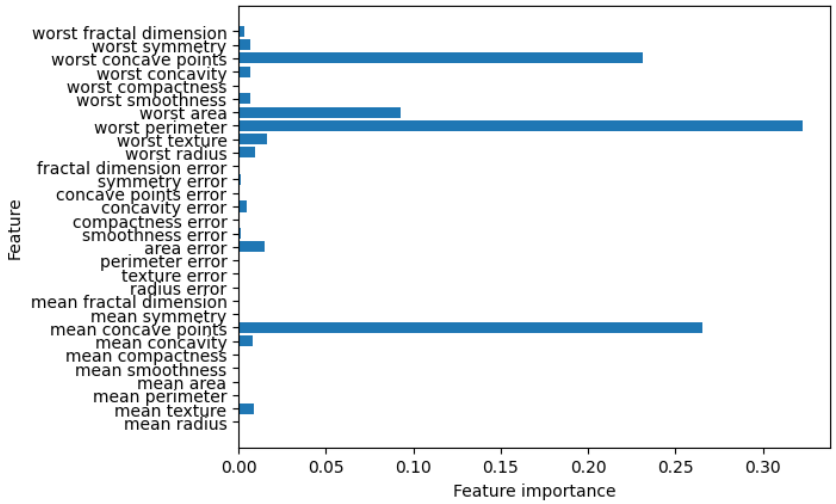

Figure: eature importances computed from a gradient boosting classifier that was fit to the Breast Cancer dataset

Supervised Learning: Ensembles of Decision Trees: Gradient boosted regression trees

Strengths, weaknesses, and parameters:

- **Strengths:**
  - Effective for various tasks, including regression and classification.
  - They can capture complex relationships in data.
  - Suitable for a mix of binary and continuous features.
  - Works well without feature scaling.
- **Weaknesses:**
  - Requires careful parameter tuning, which can be time-consuming.
  - Training may take a considerable amount of time, especially with a large number of trees.
  - Tends not to perform well on high-dimensional sparse data.

Supervised Learning: Ensembles of Decision Trees: Gradient boosted regression trees

Strengths, weaknesses, and parameters:

- **Parameters:**
  - **n_estimators:** Number of trees in the ensemble. Higher values lead to more complex models but may risk overfitting.
  - **learning_rate:** Controls the correction strength of each tree. A lower value means more trees are needed for the same complexity.
  - **max_depth:** Maximum depth of each tree. Typically set low (e.g., not deeper than five splits) to control complexity.
  - **max_leaf_nodes:** Alternatively, you can use this to limit the number of leaf nodes in each tree.

Supervised Learning: Kernelized Support Vector Machines

Outline