

Contents

1	Introduction	1
1.1	Origins of ML	1
1.2	Typical Problems of ML	2
1.3	Similarity Principle (or Compactness Hypothesis)	3
1.4	Features	5
1.4.1	Table Data	5
1.4.2	Non-Table Data	5
1.5	Distances (or Metrics)	5
1.5.1	Euclidean or ℓ_2 distance	6
1.5.2	Manhattan or ℓ_1 distance	7
1.5.3	ℓ_p distance ($1 \leq p < \infty$)	8
1.5.4	Counting or Hamming distance	8
1.6	Notation	9
1.7	Quality Metrics Examples	11
1.8	Exercises	11
2	k Nearest Neighbors (kNN)	15
2.1	kNN for Classification Problem	15
2.1.1	k Nearest Neighbors (kNN): training	16
2.1.2	k Nearest Neighbors (kNN): prediction	17
2.2	Model Comparison or How to Choose k ?	18
2.2.1	K -Fold Cross-Validation	19
2.2.2	How to choose k in kNN?	20
2.2.3	Overfitting and Generalization	21
2.3	Weighted k Nearest Neighbors (kNN)	21
2.3.1	Formulas for weights	22
2.4	kNN for Regression Problem	23
2.4.1	k Nearest Neighbors (kNN) for Regression: training	24
2.4.2	k Nearest Neighbors (kNN) for Regression: prediction	24
2.4.3	k Nearest Neighbors (kNN) Weighted Regression	25
2.4.4	Pros and Cons of kNN	25
2.5	Exercises	26
3	Exploratory Data Analysis (EDA)	27
3.1	Box-Plot and Outliers	27
3.2	Binarization (or Binning)	28

Contents

3.3	Feature/Target Distribution	29
3.3.1	Categorical (Discrete)	29
3.3.2	Numerical (Continuous)	29
3.4	Feature Importance	31
3.4.1	One-Factor Analysis	31
3.4.2	Two-Factor Analysis	31
3.4.3	Correlation	31
3.4.4	Mutual Information	31
3.4.5	By Model	31
3.5	Feature Creation	31
3.5.1	Scaling and Functional Transformations	31
3.5.2	Binning	31
3.5.3	One-Hot Encoding for Nominal Features	31
3.5.4	Dealing with Missing Values	31
3.5.5	Combining Features	31
3.5.6	Dimensionality Reduction	31
3.6	Transformations of the Target	31
4	Linear Regression	33
4.1	Optimization of Smooth Convex Functions	34
4.2	Gradient Decent in 1D	35
4.3	Optimization and Gradient Decent in 2D	36
4.3.1	Partial Derivatives and Gradient Vector	36
4.3.2	Minimization of a Smooth Convex Function in 2D	38
4.3.3	Gradient Descent Algorithm	39
4.4	Least Mean Squares for Linear Regression	40
4.4.1	Simple Linear Regression	40
4.5	Closed Form Solution for Linear Regression	41
4.6	Regression Metrics	42
4.7	Statistical Linear Regression	43
4.8	Exercises	43
5	Stochastic Gradient Descent	47
5.1	SGD	47
5.2	Mini-Batch SGD	48
5.3	Early Stopping	49
5.4	Regularization	49
5.5	Optimization for Ridge Regression	50
5.6	Optimization for Lasso Regression	51
5.6.1	Proximal Gradient Descent	51
5.7	Excercises	52

6 Model Selection	55
6.1 Feature Selection	55
6.1.1 By Criterion	55
6.1.2 By Model	56
6.1.3 By (Recursive) Elimination	57
7 Linear Models for Binary Classification	59
7.1 Geometry of Linear Classifier	60
7.2 Margin	61
7.3 Support Vector Machine (SVM)	62
7.3.1 Separable Case	62
7.3.2 Non-Separable Case	64
7.3.3 Karush — Kuhn — Tucker Theorem	65
7.3.4 Minimization with a Hinge Loss	66
7.4 F. Rosenblatt's Perceptron (1962)	68
7.4.1 Perceptron Algorithm	68
7.5 Logistic Regression	70
7.5.1 Sigmoid Function	70
7.5.2 Probabilistic Model	70
7.5.3 Maximum Likelihood	71
7.5.4 Other Metrics for Logistic Regression: Deviance, AIC, BIC	72
7.6 Classification Metrics	72
7.6.1 Confusion Matrix	73
7.6.2 <i>F</i> -score	74
7.7 Model Selection for Classification	74
7.7.1 Precision — Recall (PR) Curve	74
7.7.2 Receiver Operating Characteristic (ROC) Curve	75
7.7.3 ROC-AUC	75
7.7.4 Quality Improvement Reporting	76
7.8 Exercises	76
8 Multiclass Classification	81
8.1 One-VS-All and All-VS-All	81
8.1.1 One - VS - all	81
8.1.2 All - VS - all	82
8.2 Micro- and Macro-Averaging	82
8.3 Multiple Logistic Regression	83
8.3.1 Cross-Entropy Loss	83
8.4 Bayesian Classification	84
8.4.1 Naïve Bayes	84
9 Decision Trees	87
9.1 Geometry of Decision Tree Classifier	88

Contents

9.2 Leaf's Values	90
9.2.1 How to split one node? Classification	91
9.2.2 Impurity	91
9.2.3 Impurity	93
9.2.4 Gini VS Entropy	94
9.3 Impurity for Regression	94
9.3.1 Decision Tree Formula	94
10 Models Ensembles	95
10.1 Voting	95
10.2 Bootstrap Aggregation (Bagging)	95
10.3 Random Forest	95
10.3.1 Feature Importance	95
10.4 Boosting	95
10.4.1 Gradient Boosting Machine (GBM)	95
10.5 Stacking and Blending	95
11 Clustering	97
11.1 k -Means and k -Medoids	97
11.2 Hierarchical Clustering	97
11.3 Mixture of Gaussians	97
12 Dimensionality Reduction	99
12.1 PCA	99
12.1.1 Problem Formulation	99
12.1.2 Statistical Interpretation	102
12.1.3 Geometric Interpretation	103
12.1.4 Connection with Singular Value Decomposition (SVD)	106
12.2 t-Distributed Stochastic Neighbor Embedding (t-SNE)	106
12.2.1 One Degree of Freedom in t-Distribution	106
12.2.2 Other Degrees of Freedom	109
12.3 Orthogonal Procrustes Problem	111

1 Introduction

1.1 Origins of ML

During World War II, in 1944, the Germans constructed flying bombs and rockets to attack Great Britain. No doubt these weapons created a lot of damage and destroyed many lives.



Figure 1.1: V-1 flying bomb¹

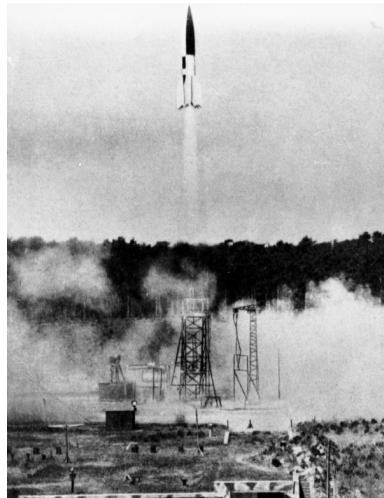


Figure 1.2: V-2 rocket²

The V-2 weapons was impossible to bring down, they were the world's first long-range guided ballistic missiles, but the V-1 weapons, if they were detected, could be defeated by anti-aircraft guns or fighter aircraft. MIT Radiation Laboratory has built a very accurate radar (see Fig. 1.4) that helped destroy 80% of the V-1 missiles. Norbert Wiener was working in this laboratory on the automatic aiming of the anti-aircraft guns. This can be considered as the first machine learning task: given the readings from the radar, calculate the trajectory of the gun to hit the V-1 weapon.

N. Wiener coined the term *cybernetics* in his 1948 book *Cybernetics: Or Control and Communication in the Animal and the Machine*. In 1950s the term *artificial intelligence*

¹By Bundesarchiv, Bild 146-1973-029A-24A / Lysiak / CC-BY-SA 3.0, CC BY-SA 3.0 de, <https://commons.wikimedia.org/w/index.php?curid=5482719> A video of its operation https://youtu.be/_i5Vq3wYPXU

²By Bundesarchiv, Bild 141-1880 / CC-BY-SA 3.0, CC BY-SA 3.0 de, <https://commons.wikimedia.org/w/index.php?curid=5338474> A video of its operation <https://youtu.be/amL3CnZNCb0>

1 Introduction

became more popular, it rooted in the works of Alan Turing in formal languages, which is related to the formal logic and automated proofs.

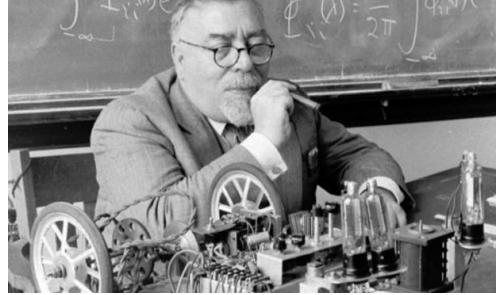


Figure 1.3: Norbert Wiener³

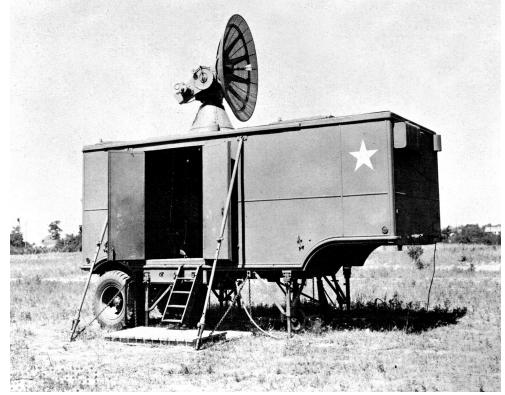


Figure 1.4: The SCR-584 system⁴

In 1974, because of a low outcome, the governments cut off the research grants in this area, this period is known as *AI winter*. It seems that using less general term *machine learning* it was easier to get funds for the research. On the other hand, in 1990s new applications of neural network models appeared with increasing computational powers.

1.2 Typical Problems of ML

1. Consider the following problem: transform hours into minutes. Everyone will solve it using the formula

$$f(x) = 60x,$$

where x is hours.

2. Now, let's think about another problem. Given the following review:

“Well, that movie was just fantastic. I couldn't stop laughing the entire time. The plot was so riveting, and the acting was top-notch. Definitely worth my time. Not.”

you should answer the question: is this review

- Positive
- Neutral
- Negative

³<https://www.arshake.com/en/norbert-wiener/>

⁴By Bundesarchiv, Bild 141-1880 / CC-BY-SA 3.0, CC BY-SA 3.0 de, <https://commons.wikimedia.org/w/index.php?curid=5338474>

1.3 Similarity Principle (or Compactness Hypothesis)

This task is known as *sentiment analysis*.

By analogy with the previous problem, one may want to formalize it as follows. Given a text x , construct a function that returns three values:

$$f(x) = 1, 0, \text{ or } -1.$$

But what does $f(x)$ mean? How can one evaluate a function of a text?

This is a typical setup for machine learning.

- The input and output of the algorithm don't have clear dependencies.
- The solution can't be represented by an exact formula.
- At the same time, we have some examples.
- We are OK with an approximate solution.
- We can use the examples for predictions!

1.3 Similarity Principle (or Compactness Hypothesis)

We'll start with an example. Look at the pictures of two trees below. In Fig. 1.5, there is a fir tree, while in Fig. 1.6, there is a pine tree.



Figure 1.5: Fir Tree⁵



Figure 1.6: Pine Tree⁶

1 Introduction

You can notice some similarities and some differences between these two images.

For the fir: the trunk is not visible, the greenery is denser and bright green color. For the pine: visible trunk and dull color. Both trees have upward trunks.

Now, you are given a new image in Fig. 1.7 and you should answer the question if it's a fir or a pine?



Figure 1.7: Fir or Pine Tree?⁷

To answer this question, you can list the characteristics of the new image and compare them with two previous images. On the new image, the branches are upwards, the trunk is not visible, the greenery is dense, and the color is light. You see that it has more common features with the tree in Fig. 1.5. Therefore, your answer will be that there is a fir in Fig. 1.7.

It looks very natural to group similar objects together.

Proposition 1.1

Similarity Principle: We will assume that if the features are similar, then the objects are from the same class.

⁵Photo by David Iliff https://commons.wikimedia.org/wiki/File:Abies_pinsapo_var._tazaotana,_Wakehurst_Place,_UK_-_Diliff.jpg

⁶<https://www.natesnursery.com/trees-of-the-high-desert-eldarica-pine/>

⁷<https://matterntrees.com/product/concolor-fir-cut>

In the remaining part of this chapter, we will formalize the words used above: features and similarity.

1.4 Features

1.4.1 Table Data

- **Binary/Boolean {0, 1}** If we look at a person's profile and are interested in whether they have a college degree or not, this is an example of a binary feature. Of course, this information can be not available, but we will talk about missing data later.
- **Numerical** For example, the size of an apartment.
- **Categorical (ordinal and nominal)** If we buy a car, usually, it has a choice of different colors. Another example of categorical feature is the trim of the car. The trim is an **ordinal** feature, because it can be ordered, while the color is a **nominal** feature, because we can't order the cars by color.
- **Datetime**
- **Geospatial**

1.4.2 Non-Table Data

- **Textual**
- **Image**
- **Audio**

1.5 Distances (or Metrics)

To measure the closeness of two features we can use the notion of distance.

Definition 1.1

A function

$$d : \mathbb{X} \times \mathbb{X} \rightarrow [0, +\infty) \quad (1.1)$$

is said to be a **distance** (or **metric**), if the following conditions are satisfied:

1. $d(x^{(1)}, x^{(2)}) = 0 \Leftrightarrow x^{(1)} = x^{(2)}$,
2. $d(x^{(1)}, x^{(2)}) = d(x^{(2)}, x^{(1)})$ for all $x^{(1)}, x^{(2)} \in \mathbb{X}$ (symmetry)
3. $d(x^{(1)}, x^{(2)}) \leq d(x^{(1)}, x^{(3)}) + d(x^{(2)}, x^{(3)})$ for all $x^{(1)}, x^{(2)}, x^{(3)} \in \mathbb{X}$ (the triangle inequality)

Remark 1.1

In mathematics, there is also a notion of a *norm* of the vector x , denoted by $\|x\|$. By definition, a **norm** is a non-negative function

$$\|\cdot\|: \mathbb{X} \rightarrow [0, \infty), \quad (1.2)$$

such that

1. $\|x\| = 0 \Leftrightarrow x = 0$.
2. $\|\lambda x\| = |\lambda| \|x\|$ for any number λ and $x \in \mathbb{X}$.
3. $\|x^{(1)} + x^{(2)}\| \leq \|x^{(1)}\| + \|x^{(2)}\|$ for all $x^{(1)}, x^{(2)} \in \mathbb{X}$.

A distance can always be defined by a norm as follows:

$$d(x^{(1)}, x^{(2)}) = \|x^{(1)} - x^{(2)}\|. \quad (1.3)$$

We will often use the second notation for the distance. The norm of a vector represents its length, or distance from the origin:

$$d(0, x) = \|0 - x\| = \|x\|. \quad (1.4)$$

Sometimes, we are able to define a distance, but not a norm (like in case of Hamming distance).

1.5.1 Euclidean or ℓ_2 distance

This is a distance familiar to us from the school:

$$d(x^{(1)}, x^{(2)}) = \|x^{(1)} - x^{(2)}\|_2 = \sqrt{\sum_{j=1}^d (x_j^{(1)} - x_j^{(2)})^2}. \quad (1.5)$$

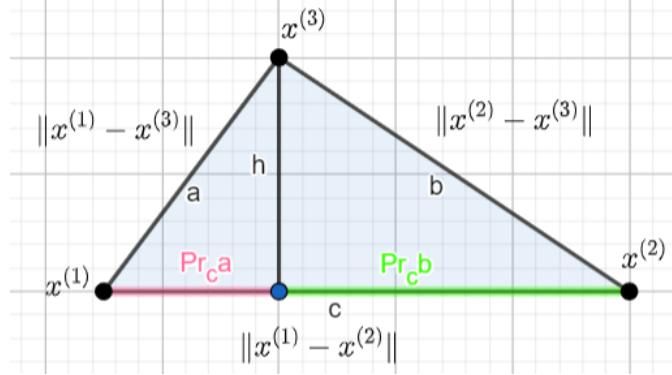
Let's check that it satisfies Definition 1.1. Obviously, it is non-negative.

1. If $d(x^{(1)}, x^{(2)}) = 0$, it means that

$$(x_1^{(1)} - x_1^{(2)})^2 + (x_2^{(1)} - x_2^{(2)})^2 + \dots + (x_d^{(1)} - x_d^{(2)})^2 = 0.$$

The sum of non-negative terms can be zero only if each term is equal to zero. Therefore, $x_j^{(1)} = x_j^{(2)}$ for every $j = 1, \dots, d$. Opposite is straightforward: from $x^{(1)} = x^{(2)}$ it follows that $x_j^{(1)} = x_j^{(2)}$ and $d(x^{(1)}, x^{(2)}) = 0$.

2. The symmetry follows from the equality $(a - b)^2 = (b - a)^2$.
3. In case of \mathbb{R}^2 , this inequality follows from an observation that the triangle sides are not shorter than their projections

**Figure 1.8:** $a + b > c$

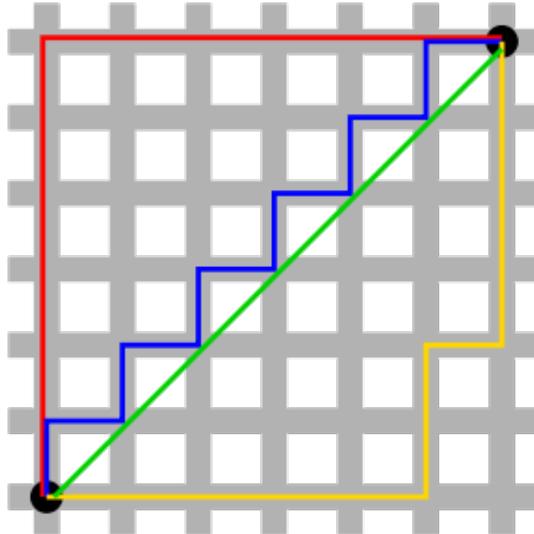
The case $d > 2$ is covered in the Exercises and in ℓ_p distance.

1.5.2 Manhattan or ℓ_1 distance

$$d(x^{(1)}, x^{(2)}) = \|x^{(1)} - x^{(2)}\|_1 = \sum_{j=1}^d |x_j^{(1)} - x_j^{(2)}| \quad (1.6)$$

The properties of the distance are easy to check. The triangle inequality follows from the inequality

$$|a + b| < |a| + |b|. \quad (1.7)$$

**Figure 1.9:** Manhattan distance in \mathbb{R}^2 ⁸

⁸https://en.wikipedia.org/wiki/Taxicab_geometry

1.5.3 ℓ_p distance ($1 \leq p < \infty$)

$$d(x^{(1)}, x^{(2)}) = \|x^{(1)} - x^{(2)}\|_p = \sqrt[p]{\sum_{j=1}^d |x_j^{(1)} - x_j^{(2)}|^p} \quad (1.8)$$

The first two properties of the distance are similar to ℓ_2 distance. The triangle inequality for this case is known as Minkowski's inequality (we don't prove it):

$$\left(\sum_{j=1}^d |a_j + b_j|^p \right)^{1/p} \leq \left(\sum_{j=1}^d |a_j|^p \right)^{1/p} + \left(\sum_{j=1}^d |b_j|^p \right)^{1/p}. \quad (1.9)$$

Remark 1.2

It is not quite accurate to call the above metric by ℓ_p metric, but it's a standard in ML. In mathematics, ℓ_p metric is used for infinite sequences $x = (x_1, x_2, \dots)$:

$$\|x\|_p = \left(\sum_{j=1}^{\infty} |x_j|^p \right)^{1/p}. \quad (1.10)$$

1.5.4 Counting or Hamming distance

$$d(x^{(1)}, x^{(2)}) = [x^{(1)} \neq x^{(2)}] = \begin{cases} 0, & x^{(1)} = x^{(2)}, \\ 1, & x^{(1)} \neq x^{(2)}. \end{cases} \quad (1.11)$$

Here we used so called **Iverson bracket** notation

$$[True] = 1, \quad [False] = 0. \quad (1.12)$$

It is easy to check that the counting distance satisfies the distance properties. It is non-negative, it equals to zero if and only if $x^{(1)} = x^{(2)}$, it is symmetric, and to prove the triangle inequality

$$[x^{(1)} \neq x^{(2)}] \leq [x^{(1)} \neq x^{(3)}] + [x^{(3)} \neq x^{(2)}] \quad (1.13)$$

it is enough to notice that the left-hand side takes values 0 or 1, while the right-hand side can take values 0, 1, or 2.

This distance can be used for non-numeric features as well.

1.6 Notation

Definition 1.2

By x we denote **example/sample/features** of the data. If it has a numerical representation, e.g., pixels in an image, it has several components: $x = (x_1, x_2, \dots, x_d)$. From probabilistic point of view it is considered as a random vector X on \mathbb{R}^d .

By \mathbb{X} we denote the **sample space**, all possible examples of data.

By y we denote the response of the system usually referenced as **target**.

By \mathbb{Y} we denote the **target space**.

Typically, we have several examples of the data, e.g., we have a database of houses for sale. Then for an individual record we'll use a superscript.

Definition 1.3

The **training data** we'll denote by ordered pairs $(x^{(i)}, y^{(i)})$, $i = 1, 2, \dots, N$.

This notation is convenient for multidimensional data: $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \dots, x_d^{(i)})$.

Definition 1.4

The goal of **ML model/algorithm** a is for given characteristics/features of the data predict the target:

$$a(x) \approx y. \quad (1.14)$$

Example. We want to predict the price of an apartment given its total area x in square feet. The simplest model is the linear model:

$$a(x) = w_0 + w_1 x. \quad (1.15)$$

To build the model (1.15) we need to determine (or to learn) the coefficients w_0 and w_1 .

Definition 1.5

The process of identifying of the coefficients (weights in front the features) from the available data is called **training/learning** of the model.

Definition 1.6

If all targets $y^{(i)}$ are known, we will talk about **supervised learning** task.

If $y^{(i)}$ are known only for a subset of indices i , then we will call this task **semi-supervised learning**.

In case none of $y^{(i)}$ are known, we will talk about **unsupervised learning**.

1 Introduction

When the model is trained, it can be used for prediction on the unseen data. Let's say we trained our linear model (1.15) and found $w_0 = 100,000$ and $w_1 = 0.1$, then for an apartment with area 2,000 sq ft the predicted price is $a(2,000) = 100,000 + 0.1 \times 2.000 = 300,000$.

Definition 1.7

Prediction of the model for a given input x we denote with hat:

$$\hat{y} = a(x) \quad \text{or} \quad y_{pred} = a(x). \quad (1.16)$$

To assess the quality of prediction, we should be able to compare the predicted value with the true value. Let's say, the true price of the apartment above was \$210,000. By overestimating the price as \$300,000 we could lose the opportunity to buy a good apartment. To evaluate the algorithm, the notion of loss function is used.

Definition 1.8

The **loss function** is a function of two arguments

$$Loss: \mathbb{Y} \times \mathbb{Y} \rightarrow [0, \infty), \quad (1.17)$$

such that $Loss(\hat{y}, y) = 0$ if and only if $\hat{y} = y$.

For example, if $Loss(\hat{y}, y) = |\hat{y} - y|$, then in our example $Loss(300,000, 210,000) = 90,000$.

Definition 1.9

If we consider the input data as a random vector X , then the output of the model is also a random vector $\hat{Y} = a(X)$ and one can talk about **expected loss**:

$$E[Loss(\hat{Y}, Y)] = E[Loss(a(X), Y)] = \int_{\mathbb{X} \times \mathbb{Y}} Loss(a(x), y) p(x, y) dx dy. \quad (1.18)$$

Usually, the distribution of the data (X, Y) is unknown and empirical distribution is used, i.e., a discrete distribution that uniformly assigns probability $1/N$ to every sample.

Definition 1.10

By **empirical loss** (or **empirical risk**) we mean the expected loss with empirical probability density function:

$$E[Loss(a(X), Y)] \approx \frac{1}{N} \sum_{i=1}^N Loss(a(x^{(i)}), y^{(i)}). \quad (1.19)$$

The typical loss functions are

1. 0-1 loss (counting or Hamming distance)

$$Loss(a(x), y) = [a(x) \neq y]$$

2. ℓ_2 loss (Euclidean distance)

$$Loss(a(x), y) = \sqrt{(a_1 - y_1)^2 + \dots + (a_d - y_d)^2}$$

3. ℓ_1 loss (Manhattan distance)

$$Loss(a(x), y) = |a_1 - y_1| + \dots + |a_d - y_d|$$

1.7 Quality Metrics Examples

After the model was trained, we can test it on some set of data (typically on the new, unseen during training data), and asses the quality of the model with the same as loss or another metric.

Example 1.

$$MSE = \frac{1}{N} \sum_{i=1}^N \left(a(x^{(i)}) - y^{(i)} \right)^2$$

Example 2.

$$Error = \frac{1}{N} \sum_{i=1}^N \left[a(x^{(i)}) \neq y^{(i)} \right] = \frac{\#\text{incorrect predictions}}{N}$$

$$Accuracy = \frac{1}{N} \sum_{i=1}^N \left[a(x^{(i)}) = y^{(i)} \right] = \frac{\#\text{correct predictions}}{N}$$

1.8 Exercises

Problem 1. Assume, you work with a taxi system. The dataset consists of various columns that capture different aspects of the clients. Classify every column type (binary, numerical, categorical: ordinal and nominal, textual, datetime, geospatial, audio, etc.)

How many trips during a day	Average expenses on taxi	How often premium was used	Age	Agreed to upgrade to premium
1	6	15	27	yes
2	7	2	54	no
...

1 Introduction

Problem 2. Consider the data about a new user with the following information available:

How many trips during a day	Average expenses on taxi	How often premium was used	Age	Agreed to upgrade to premium
3	10	0	25	

Based on the *compactness principle*, use ℓ_1 -distance to predict if this user is likely to upgrade the trip. In other words, calculate the distance between these data and each record in the table from Problem 1. Based on the shortest distance, predict whether the user is likely to agree to an upgrade.

Problem 3. Consider now slightly different data (in one cell):

How many trips during a day	Average expenses on taxi	How often premium was used	Age	Agreed to upgrade to premium
1	6	100	27	yes
2	7	2	54	no
...

How the prediction for the passenger from Problem 3 will be changed? Does it seem reasonable? If not, how could you change the distance function to get more reasonable prediction?

Problem 4. Prove the triangle inequality for Euclidean distance.

Solution. To prove the triangle inequality let's look at the following sum:

$$\begin{aligned} \sum_{j=1}^d \sum_{k=1}^d (a_j b_k - b_j a_k)^2 &= \sum_{j=1}^d \sum_{k=1}^d (a_j^2 b_k^2 - 2a_j a_k b_j b_k + b_j^2 a_k^2) \\ &= 2 \sum_{j=1}^d \sum_{k=1}^d a_j^2 b_k^2 - 2 \sum_{j=1}^d \sum_{k=1}^d a_j a_k b_j b_k \end{aligned} \tag{1.20}$$

On the other hand,

$$\begin{aligned} \left(\sum_{j=1}^d a_j b_j \right)^2 &= \left(a_1 b_1 + \sum_{j=2}^d a_j b_j \right)^2 = a_1^2 b_1^2 + 2a_1 b_1 \sum_{j=2}^d a_j b_j + \left(\sum_{j=2}^d a_j b_j \right)^2 \\ &= \sum_{j=1}^d a_j^2 b_j^2 + 2 \sum_{j < k} a_j b_j a_k b_k = \sum_{j=1}^d \sum_{k=1}^d a_j b_j a_k b_k \\ &= \sum_{j=1}^d \sum_{k=1}^d a_j^2 b_k^2 - \frac{1}{2} \sum_{j=1}^d \sum_{k=1}^d (a_j b_k - b_j a_k)^2 \leq \left(\sum_{j=1}^d a_j^2 \right) \left(\sum_{k=1}^d b_k^2 \right). \end{aligned} \tag{1.21}$$

From the last inequality, which is known as Cauchy—Schwarz inequality, it follows that

$$\begin{aligned}
 \|x_j^{(1)} - x_j^{(2)}\|_2^2 &= \sum_{j=1}^d (x_j^{(1)} - x_j^{(2)})^2 = \sum_{j=1}^d (x_j^{(1)} - x_j^{(3)} + x_j^{(3)} - x_j^{(2)})^2 \\
 &= \sum_{j=1}^d ((x_j^{(1)} - x_j^{(3)})^2 + 2(x_j^{(1)} - x_j^{(3)})(x_j^{(3)} - x_j^{(2)}) + (x_j^{(3)} - x_j^{(2)})^2) \\
 &\leq \sum_{j=1}^d (x_j^{(1)} - x_j^{(3)})^2 + 2\sqrt{\sum_{j=1}^d (x_j^{(1)} - x_j^{(3)})^2 \sum_{j=1}^d (x_j^{(3)} - x_j^{(2)})^2} + \sum_{j=1}^d (x_j^{(3)} - x_j^{(2)})^2 \\
 &= \left(\sqrt{\sum_{j=1}^d (x_j^{(3)} - x_j^{(2)})^2} + \sqrt{\sum_{j=1}^d (x_j^{(3)} - x_j^{(2)})^2} \right)^2 = (\|x_j^{(1)} - x_j^{(3)}\|_2 + \|x_j^{(3)} - x_j^{(2)}\|_2)^2.
 \end{aligned}$$

Problem 5. By an open circle (ball) of radius r with the center at a we mean the set of points $x \in \mathbb{R}^2$, such that $d(a, x) < r$:

$$B_r(a) = \{x \in \mathbb{R}^2 : d(a, x) < r\}. \quad (1.22)$$

Draw a unit circle centered at the origin using ℓ_2 , ℓ_1 , and counting distances.

2 k Nearest Neighbors (kNN)

2.1 kNN for Classification Problem

Definition 2.1

In the **classification** task, the target space is finite:

$$\mathbb{Y} = \{1, 2, \dots, C\} \quad \text{or} \quad \mathbb{Y} = \{0, 1, \dots, C - 1\}, \quad \text{or} \quad \mathbb{Y} = \{-1, 1\}. \quad (2.1)$$

In the **regression** task, the target can take any real number: $\mathbb{Y} = \mathbb{R}$.

Remark 2.1

There is a **multi-label classification**, when multiple labels can be assigned to one object $\mathbb{Y} = \{0, 1\}^C$, and **multiple regression**, when $\mathbb{Y} = \mathbb{R}^n$.

In Fig. 2.1, the data with two-dimensional feature space and a binary target is presented. There are three points: red, blue, and black. How would you classify them?

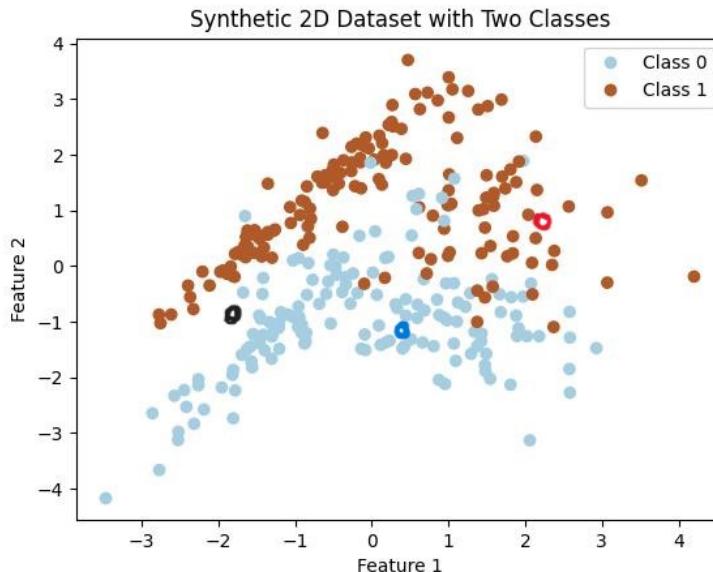


Figure 2.1: Data of two classes and three points to classify.

2 k Nearest Neighbors (kNN)

Most likely, you would classify the red point as Class 1, the blue point as Class 0, but you would have some difficulty to classify the black point.

This intuitive idea can be formalized in terms of neighboring point.

Definition 2.2

By a **neighbourhood** of a point x we will mean a ball of some radius r centered at the point x :

$$B_r(a) = \{x \in \mathbb{X} : d(a, x) < r\}. \quad (2.2)$$

The points inside the neighbourhood are called **neighbours**.

In the figure¹ below, the solid ball contains three neighbours and the dashed ball contains five neighbours.

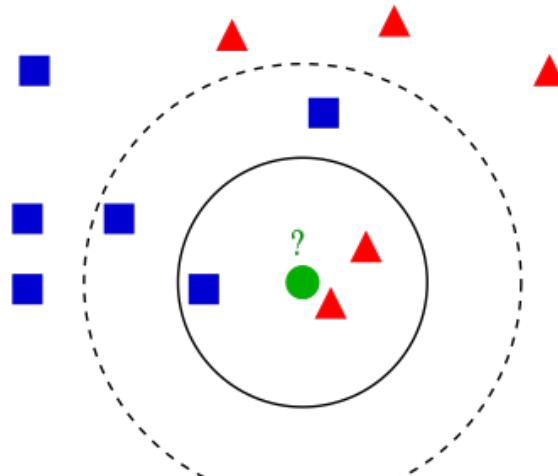


Figure 2.2

If we use three closest neighbours for the classification, then we would classify the green point as a triangle, because we have more triangles in the neighbourhood.

This example brings us to the k nearest neighbors method.

2.1.1 k Nearest Neighbors (kNN): training

- We are given $(x^{(i)}, y^{(i)})$, $i = 1, 2, \dots, N$.
- $\mathbb{Y} = \{1, 2, \dots, C\}$ (classification problem).
- training = memorizing of the given data.

¹https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

2.1.2 k Nearest Neighbors (kNN): prediction

- We have a new feature x .
- Determine the distances between the new feature and $x^{(i)}$: $d(x, x^{(i)})$.
- Rearrange the objects by the closeness to x :

$$d(x, x^{(i_1)}) \leq d(x, x^{(i_2)}) \leq \dots \leq d(x, x^{(i_N)}). \quad (2.3)$$

- Look at the first k labels and assign the class with the highest number of representatives:

$$a(x) = \operatorname{argmax}_{c \in \mathbb{Y}} \sum_{s=1}^k [y^{(i_s)} = c]. \quad (2.4)$$

Let's illustrate this algorithm with $k = 3$ on the example in Fig. 2.2. Assume, the points were indexed somehow.

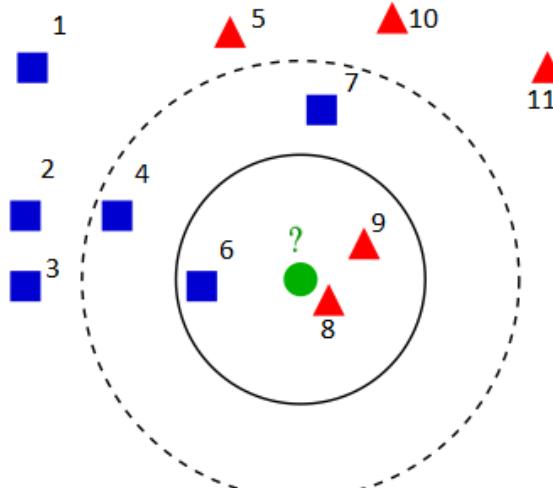


Figure 2.3

Reindex them in the increasing order of ℓ_2 distances from the green point:

$$i_1 = 8, i_2 = 9, i_3 = 6, i_4 = 7, \dots$$

In this case, $\mathbb{Y} = \{\blacksquare, \blacktriangle\}$. Calculate two sums:

$$\sum_{s=1}^3 [y^{(i_s)} = \blacksquare] = [y^{(8)} = \blacksquare] + [y^{(9)} = \blacksquare] + [y^{(6)} = \blacksquare] = 1 \quad \text{and} \quad \sum_{s=1}^3 [y^{(i_s)} = \blacktriangle] = 2$$

Finally,

$$\max_{c \in \{\blacksquare, \blacktriangle\}} \sum_{s=1}^3 [y^{(i_s)} = c] = 2 \quad \text{and} \quad a(\bullet) = \operatorname{argmax}_{c \in \{\blacksquare, \blacktriangle\}} \sum_{s=1}^3 [y^{(i_s)} = c] = \blacktriangle.$$

2 k Nearest Neighbors (kNN)

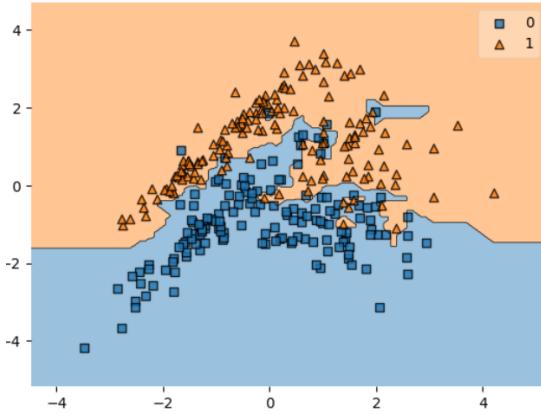


Figure 2.4: $k = 1, p = 1$

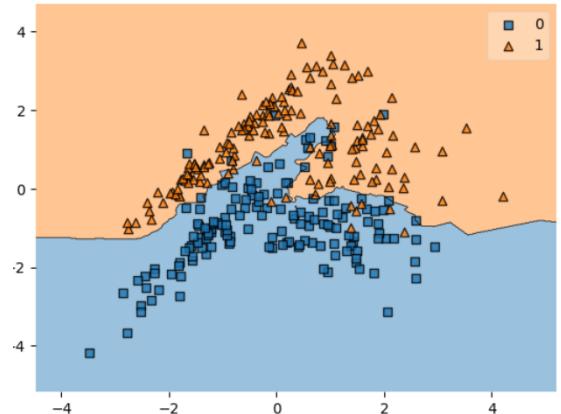


Figure 2.5: $k = 5, p = 2$

2.2 Model Comparison or How to Choose k ?

Let's apply kNN model to the data in Fig. 2.1 with ℓ_1 distance and $k = 1$ (see Fig. 2.4) and with ℓ_2 distance and $k = 5$ (see Fig. 2.5). The results of classification depend on the choice of the number of neighbors and the metric being used. These parameters are referred to as *hyper parameters*.

Definition 2.3

The parameters of the model that cannot be tuned during training are called **hyper parameters**.

We cannot tune k or the choice of distance during the training, these parameters correspond to the current model. But we can use different (hyper) parameters and train the model again, this will be a different model. Comparing somehow those two models, we can choose one that performs better, and this is how we can choose the hyper parameters.

The main method to evaluate the model's performance is to look at the quality metric on the new data. In practice, we don't want to collect new data, and we split the existing data into two folds: one we use for training and the other one is used for the final report.



Figure 2.6: Train/Test Split

2.2 Model Comparison or How to Choose k ?

If we want to tune hyper parameters, we can use one of the following approaches²:

- Evaluate model on the validation set
- Use cross-validation

To tune the hyper parameters and choose one model, we can

1. Train different models with different hyper parameters.
2. Test them on Validation (or Developing) set and choose the best model.
3. Using the best model, report the final metric on the Testing set.



Figure 2.7: Train/Test Split

2.2.1 K -Fold Cross-Validation

The other approach is to split training set into K equal parts and use each part as the validation set.

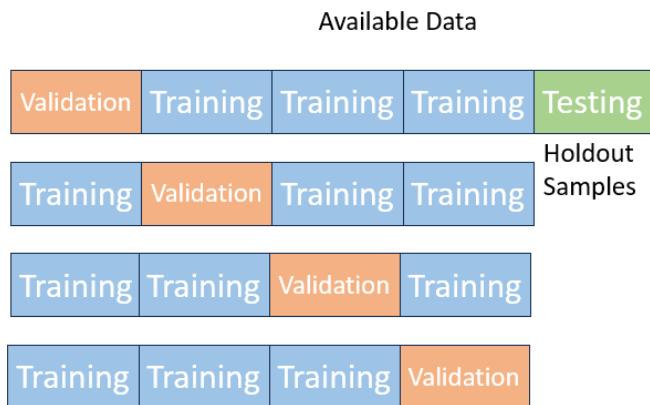


Figure 2.8: Train/Test Split

²There are other ways to compare the models, e.g., maximizing likelihood in probabilistic models. We will talk about them later.

2 k Nearest Neighbors (k NN)

1. For each model

- Train the model on the Training data from split 1 and calculate the quality metric Q_1 on the first Validation set.
- Train the model on the Training data from split 2 and calculate the quality metric Q_2 on the second Validation set.
- ...
- Train the model on the Training data from split k and calculate the quality metric Q_K on the K -th Validation set.
- report the average quality metric over all splits.

$$Q = \frac{1}{K} \sum_{i=1}^K Q_i. \quad (2.5)$$

2. Choose the best model based on the quality metric (2.5).
3. Using the best model, report the final metric on the Testing set.

Cross-validation is used when the training set is not big enough to cut a Validation set from it. Additionally, it takes care of the uneven splits (when in Validation set similar data were sampled).

2.2.2 How to choose k in kNN?

Let's demonstrate, how to choose the best number of neighbors for a fixed distance function. With a slight abuse of the previous notation, we use the testing set as the validation set. We calculate accuracy on the Training set and on the Testing set (which

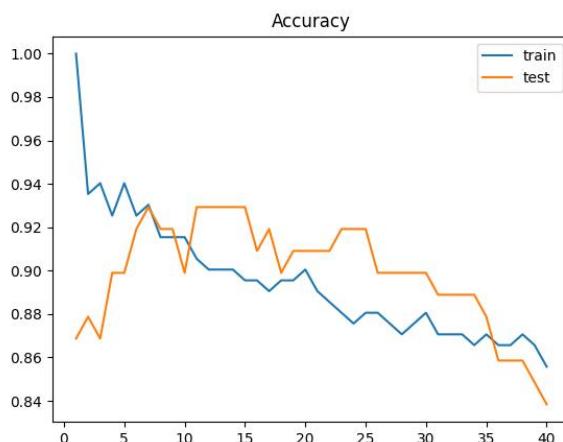


Figure 2.9: Accuracy for different value of k

2.3 Weighted k Nearest Neighbors (k NN)

should be the Validation set) for different k and plot them in one figure. The region, where these curves meet for the first time can be considered as an optimal one. In our case, the optimal $k \in [6, 9]$.

2.2.3 Overfitting and Generalization

We will try to explain, why we have chosen the number of neighbors at the point where training and validation curves met for the first time.

Let's start with an analogy. Imaging, you are preparing for an exam. You can choose two strategies: the first strategy is to memorize all the questions without understanding of the topics. It will take some time, but you will get a good grade. The other strategy is to understand the topics and instead of memorization derive the formulas from the simpler concepts. Which one would you choose?

In ML, your model can describe well the training data, but have a poor performance on the new data. This is called **memorisation** or **overfitting**. If your model performs well on the new data, then we say, it has a good **generalization** ability.

That is why we usually train the model until its performance on the unseen data begins worsen (or stops improving).

2.3 Weighted k Nearest Neighbors (k NN)

Let's look again at Fig. 2.3 with $k = 5$ neighbors. If we count the neighbors of each class

$$\sum_{s=1}^5 \left[y^{(i_s)} = \blacktriangle \right] = \begin{matrix} & \\ & 1 \\ \left[y^{(8)} = \blacktriangle \right] + \left[y^{(9)} = \blacktriangle \right] + \left[y^{(6)} = \blacktriangle \right] + \left[y^{(7)} = \blacktriangle \right] + \left[y^{(4)} = \blacktriangle \right] & = 2, \end{matrix}$$

$$\sum_{s=1}^5 \left[y^{(i_s)} = \blacksquare \right] = \begin{matrix} & \\ & 0 \\ \left[y^{(8)} = \blacksquare \right] + \left[y^{(9)} = \blacksquare \right] + \left[y^{(6)} = \blacksquare \right] + \left[y^{(7)} = \blacksquare \right] + \left[y^{(4)} = \blacksquare \right] & = 3, \end{matrix}$$

we should assign class \blacksquare to the green point. But it doesn't look correct, because the green point is much closer to the red triangles, than to the blue squares.

To take the distance into account, we introduce the coefficients (weights) decreasing with the distance, for example,

$$\sum_{s=1}^5 \left[y^{(i_s)} = \blacktriangle \right] = \frac{5}{5} \cdot 1 + \frac{4}{5} \cdot 1 + \frac{3}{5} \cdot 0 + \frac{2}{5} \cdot 0 + \frac{1}{5} \cdot 0 = \frac{9}{5},$$

$$\sum_{s=1}^5 \left[y^{(i_s)} = \blacksquare \right] = \frac{5}{5} \cdot 0 + \frac{4}{5} \cdot 0 + \frac{3}{5} \cdot 1 + \frac{2}{5} \cdot 1 + \frac{1}{5} \cdot 1 = \frac{6}{5},$$

and choose the class with smaller sum. In this case, we would choose class \blacktriangle .

The general formula for prediction takes the following form:

$$a(x) = \operatorname{argmax}_{c \in \mathbb{Y}} \sum_{s=1}^k w_s \left[y^{(i_s)} = c \right]. \quad (2.6)$$

2 k Nearest Neighbors (kNN)

2.3.1 Formulas for weights

The weights decay, we use above is given by

$$w_s = \frac{k+1-s}{k}, \quad s = 1, 2, \dots, k. \quad (2.7)$$

Another common choice of the weight is

$$w_s = q^s, \quad 0 < q < 1, \quad s = 1, 2, \dots, k. \quad (2.8)$$

A kernel

$$w_s = K\left(\frac{d(x, x^{(i_s)})}{h}\right), \quad (2.9)$$

where K is a non-increasing function defined on $[0, 1]$ or $[0, \infty)$.

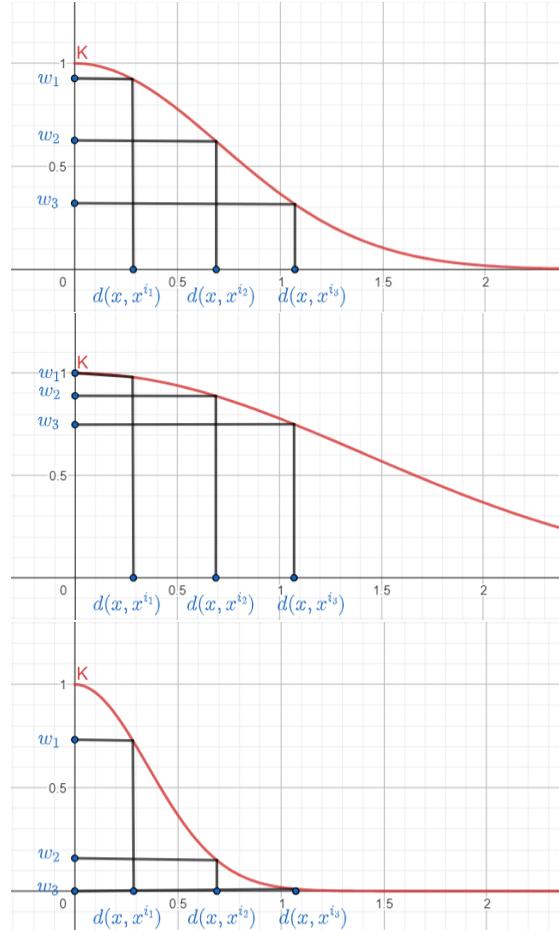


Figure 2.10: $K(d/h) = e^{-(d/h)^2}$ for $h = 1, 2$, and 0.5

With smaller parameter h we will assign smaller weights to larger distances.

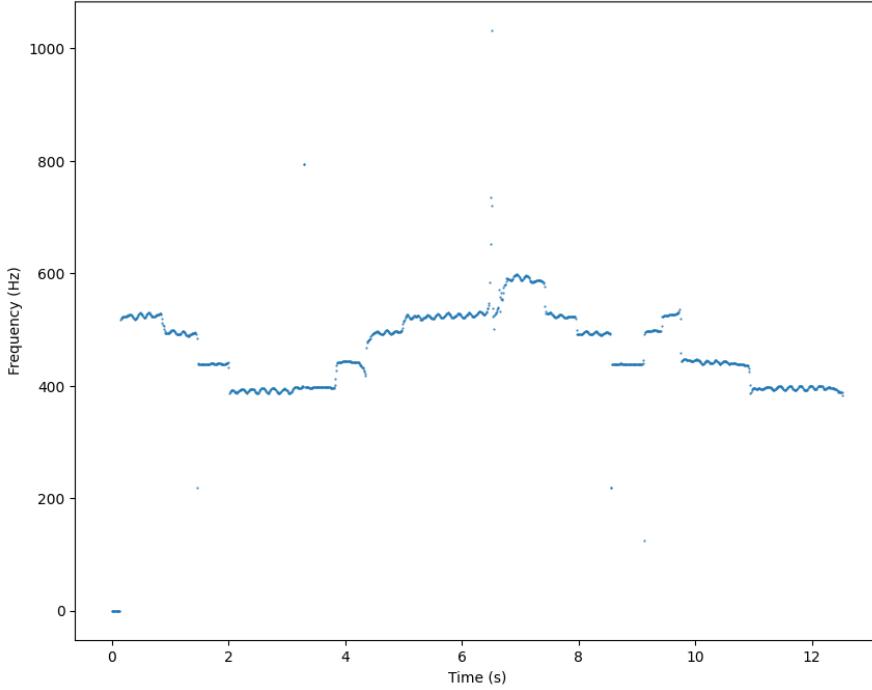


Figure 2.11: Pitch detection of a violin recording

2.4 kNN for Regression Problem

In previous section, we used neighboring data points to predict a class based on the most common labels. We want to use the information of the closer neighbors to predict continuous data. As an example, look at the output of the YIN algorithm for pitch detection³ in Fig. 2.11. The data plotted $(x^{(i)}, y^{(i)})$ represent time of the audio and corresponding pitch prediction. If we were interested in the pitch at time $x^{(i)} < x < x^{(i+1)}$, then most likely we would use the average

$$y \approx a(x) = \frac{y^{(i)} + y^{(i+1)}}{2}. \quad (2.10)$$

If we want to reduce the noise of the data, we can average over more points.⁴ This brings us to the kNN algorithm for continuous data prediction.

³http://audition.ens.fr/adc/pdf/2002_JASA_YIN.pdf

⁴This idea is based on the Law of Large Numbers, which says that the sum of i.i.d. random variables $X^{(i)}$ tends to their expected value:

$$\frac{X^{(1)} + X^{(2)} + \dots + X^{(N)}}{n} \approx E[X^{(1)}].$$

It is reasonable to assume that the noise appears independently and in average is close to zero.

2 k Nearest Neighbors (kNN)

Definition 2.4

By **regression task** in machine learning we mean a problem, which solution is an arbitrary number from some range, e.g., $\mathbb{Y} = \mathbb{R}$ in case of temperature prediction, or $\mathbb{Y} = [0, 1]$ in case we predict the probability.

Remark 2.2

Looking at formula (2.10) one may ask the question: isn't regression is the same as **interpolation**? The answer is no. In case of interpolation, we assign a value close to the near points. In case of the stock market prediction, the intermediate value can deviate a lot from the nearest values. In case of regression, we hope to find some pattern or functional dependency between features and the target. This definition of the regression problem shouldn't be confused with the **statistical regression** problem, when we assume some probability distribution behind the data.

The other difference is that in the interpolation, the resulted curve goes through the data, in case of regression, it may not go through any of the data points.

2.4.1 k Nearest Neighbors (kNN) for Regression: training

- We are given $(x^{(i)}, y^{(i)})$, $i = 1, 2, \dots, N$.
- $\mathbb{Y} = \mathbb{R}$ (regression problem).
- training = memorizing of the given data.

2.4.2 k Nearest Neighbors (kNN) for Regression: prediction

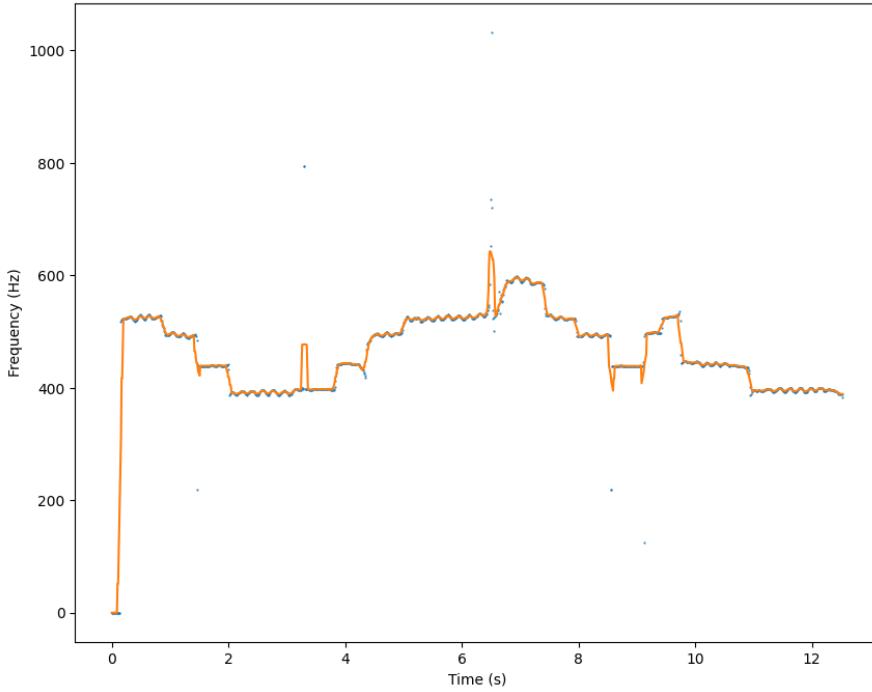
- We have a new feature x .
- Define the distance between new feature and $x^{(i)}$: $d(x, x^{(i)})$.
- Rearrange the objects by the closeness to x :

$$d(x, x^{(i_1)}) \leq d(x, x^{(i_2)}) \leq \dots \leq d(x, x^{(i_N)}).$$

- Look at the first k labels and assign the average of those:

$$a(x) = \frac{1}{k} \sum_{s=1}^k y^{(i_s)}. \quad (2.11)$$

The result of kNN regression for the data in Fig. 2.11 with $k = 15$ represented in Fig. 2.12. We see some peaks caused by outliers, we will discuss it in the next chapter. This section we conclude with weighted version of kNN regression.

**Figure 2.12:** $k = 15$

2.4.3 k Nearest Neighbors (kNN) Weighted Regression

In case of weighted kNN regression, we normalize the weighted sum:

$$a(x) = \frac{\sum_{s=1}^k w_s y^{(i_s)}}{\sum_{s=1}^k w_s}. \quad (2.12)$$

If $w_s = 1$ for all $s = 1, \dots, k$, then this formula coincides with average (2.11).

Remark 2.3

When the weights are defined by a kernel function: $w_s = K\left(\frac{d(x, x^{(i_s)})}{h}\right)$, this prediction coincides with the Nadaraya — Watson estimator in the statistical non-parametric regression.

2.4.4 Pros and Cons of kNN

Pros

- If there are many examples, the algorithm gives a good prediction.

2 k Nearest Neighbors (kNN)

- Simple training.
- Few hyperparameters.
- There exist problems, where kNN works better, e.g., texts classification with many classes.

Cons

- Usually, other models work better.
- Extensive usage of storage.
- Sorting for finding the neighbors.
- The model has limited tuning abilities.

2.5 Exercises

Problem 1. Apply the kNN method with 2 neighbors to the following data.

x	0.45	-0.1	2	0.3	-0.5	0.7	0
y	+1	-1	+1	+1	-1	-1	+1

Report the accuracy.

Hint: Order the data by x first. Consider the current point unlabeled. In cases of ambiguity choose the closest point, if the points (with different labels) are equidistant, return 0 as a refusal of classification.

3 Exploratory Data Analysis (EDA)

3.1 Box-Plot and Outliers

Fig. 2.11 represents frequencies detected by YIN algorithm, but some frequencies were detected incorrectly, like 1000 Hz. We call the data with a high deviation from the average them **outliers**.

Definition 3.1

Let

$$y^{(i_1)}, y^{(i_2)}, \dots, y^{(i_N)}$$

be a variational series, i.e., the sorted sequence of the values $y^{(i)}$ ($i = 1, \dots, N$).

For a given number $0 \leq \alpha \leq 1$ the value

$$y^{(i_s)}, \quad s = \lfloor \alpha N \rfloor \quad (3.1)$$

is called the **$\alpha \cdot 100$ -th percentile**.

25-th, 50-th, and 75-th percentiles are known as **1-st, 2-d, and 3-d quartiles**.

The 2-d quartile coincides with the **median**.

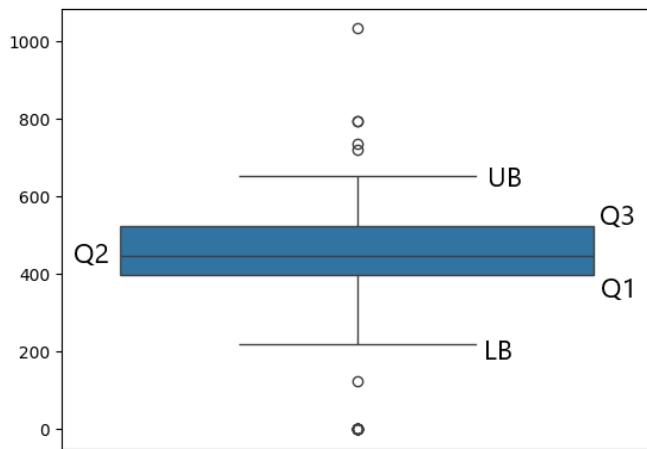


Figure 3.1: Box-plot for data in Fig. 2.11 (seaborn library)

The **box-plot diagram** represents the quartiles, there are also whiskies, we denote them by Low Bound (LB) and Upper Bound (UP). The points outside the whiskies

3 Exploratory Data Analysis (EDA)

are considered as outliers. In matplotlib and seaborn libraries, they are calculated as follows:

$$LB = Q1 - 1.5 \cdot (Q3 - Q1), \quad UB = Q3 + 1.5 \cdot (Q3 - Q1). \quad (3.2)$$

The upper and lower bounds can be defined in different ways. In case of the pitch detection, we used the 3-d and 97-th percentiles as the threshold for outliers (see Fig.3.2).

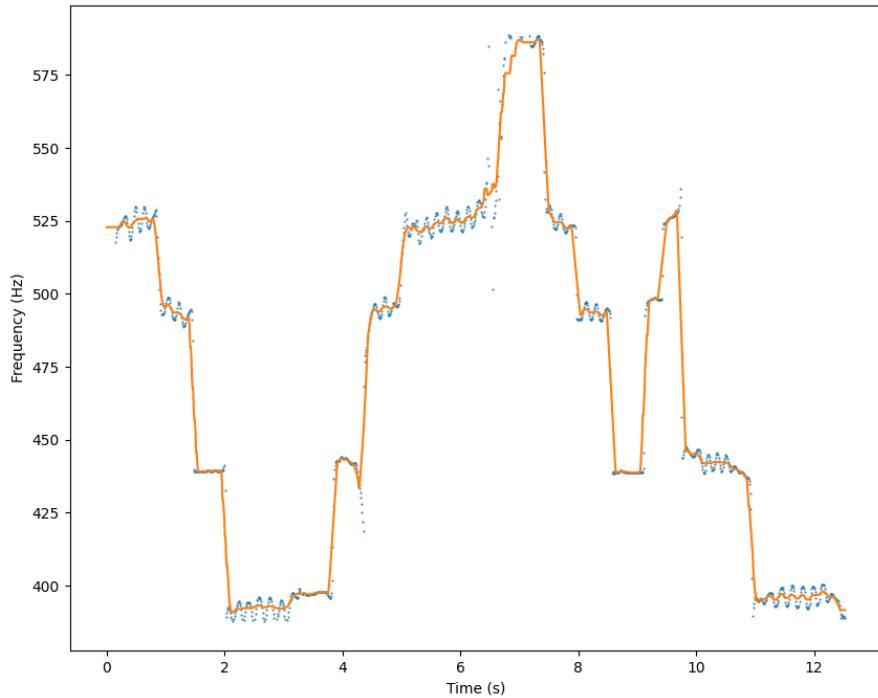


Figure 3.2: After removing points outside 3% and 97% percentiles

Remark 3.1

There is no strict definition of the outliers. Some times rare or very unusual events can be of high importance. For example, an airplane crash is a very rare event, but it can not be eliminated if we analyze the safety of the flights.

3.2 Binarization (or Binning)

In some cases, the small differences in data are not important. For example, age 60, or 61. In this case we can split the continuous feature into intervals (bins). E.g., in case of age

$$[0, 5), [5, 10), [10, 16), [16, 21), [21, 35), [35, 50), [50, \infty). \quad (3.3)$$

If we enumerate these intervals as 0, 1, ..., 6, then we can create a discrete ordinal feature from the continuous one.

3.3 Feature/Target Distribution

From Fig. 3.2 it is clear that only five notes were played. The result of binning of the detected frequencies is presented below.

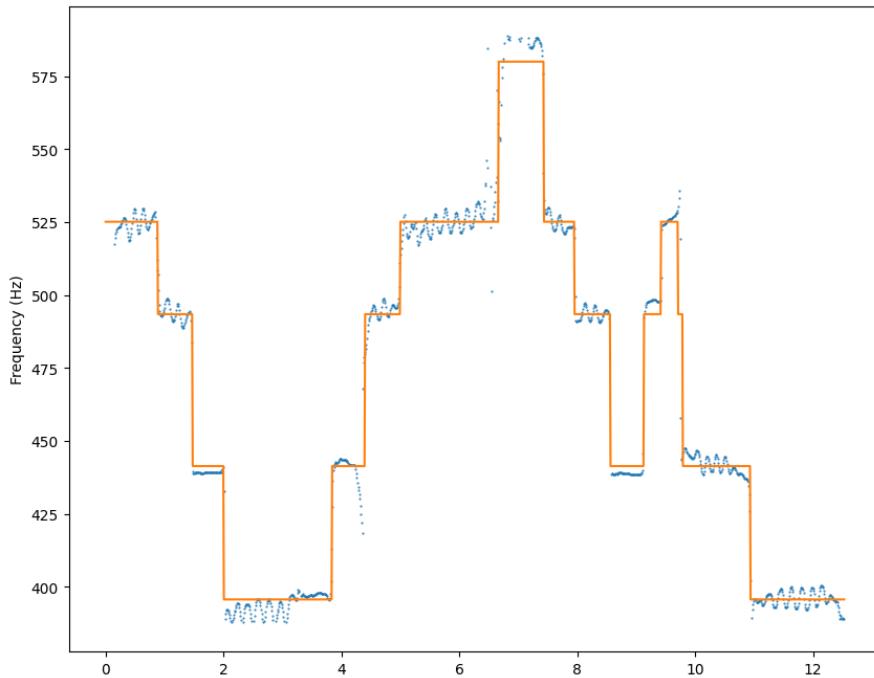


Figure 3.3: 5 bins

3.3 Feature/Target Distribution

In the remainder of this chapter, we will work with the *Titanic* dataset.¹

3.3.1 Categorical (Discrete)

To visualize a categorical feature distribution we can use **bar charts**. Fig. 3.4 shows the number of objects (Titanic's passengers) in each category of the feature *Sex*.² Fig. 3.4 shows the normalized (probability) distribution with total sum equal to one.³ The normalized bar chart can be viewed as *probability mass function* (pmf).

3.3.2 Numerical (Continuous)

The distribution of continuous data is described by the *probability density function* (pdf). In order to approximate pdf, we split the full range of the data into small intervals

¹<https://www.kaggle.com/c/titanic/>, we use its version https://raw.githubusercontent.com/iad34/seminars/master/materials/data_sem1.csv

²`sns.barplot(Data['Sex'].value_counts(dropna=False))`

³`sns.barplot(Data['Sex'].value_counts(dropna=False, normalize=True))`

3 Exploratory Data Analysis (EDA)

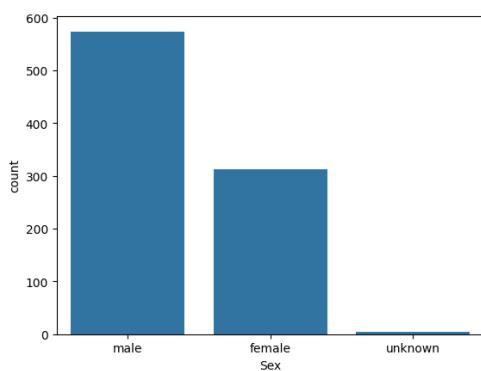


Figure 3.4: Counts (or Frequencies)

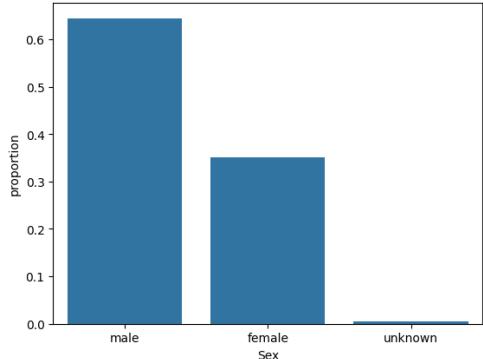


Figure 3.5: Normalized Counts (of Relative Frequencies)

(bins) and count the number of data points in these bins (see Fig. 3.6),⁴ thus we get the *histogram*. The normalized histogram (divided by the total number of samples) approximates pdf. To get a more smooth curve the *kernel density estimation* can be used.⁵

$$p(x) \approx p_h(x) = \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{x - x^{(i)}}{h}\right), \quad (3.4)$$

where K is a kernel function (see kNN regression). In Fig. 3.7

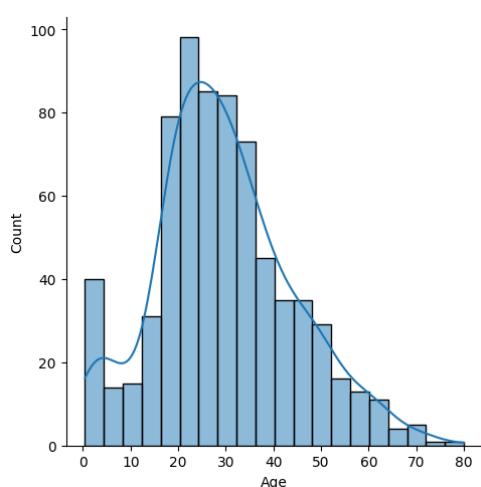


Figure 3.6: Histogram

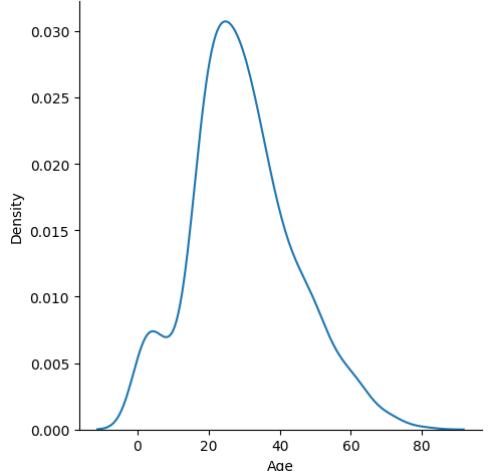


Figure 3.7: Kernel Density Estimation (over Normalized Histogram)

The following topics are discussed in the class notebook.

⁴`sns.displot(Data.Age, kde=True)`

⁵`sns.displot(Data.Age, kind='kde')`

3.4 Feature Importance

3.4.1 One-Factor Analysis

3.4.2 Two-Factor Analysis

3.4.3 Correlation

3.4.4 Mutual Information

3.4.5 By Model

3.5 Feature Creation

3.5.1 Scaling and Functional Transformations

3.5.2 Binning

3.5.3 One-Hot Encoding for Nominal Features

3.5.4 Dealing with Missing Values

3.5.5 Combining Features

3.5.6 Dimensionality Reduction

3.6 Transformations of the Target

4 Linear Regression

Consider Galton's data about the height of parents and children¹

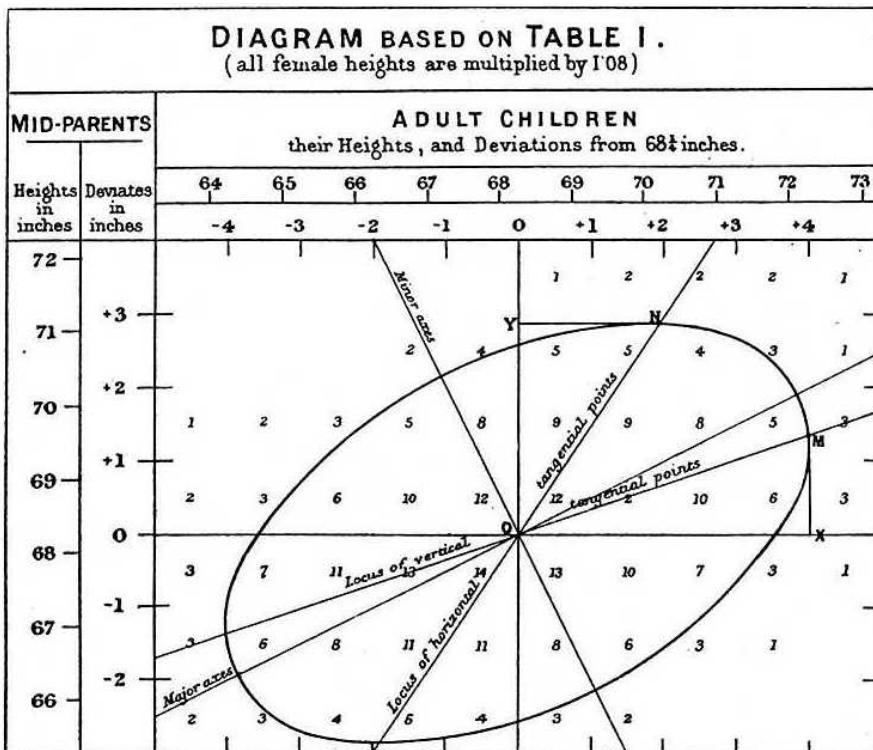


Figure 4.1: The plot from F. Galton's paper

The simplest assumption one can make: the data lie along a line. This data approximation is called *linear regression*²

F. Galton introduced the term *regression*. In Fig. 4.1, you can see that “extremely tall parents tend to have children who are taller than average and extremely small parents tend to have children who are smaller than average, but in both cases the children tend to be closer to the average than were their parents.”³

We plotted the data from the Galton dataset⁴ in Fig. 4.2

¹Francis Galton. Regression towards mediocrity in hereditary stature. The Journal of the Anthropological Institute of Great Britain and Ireland. 15: 246–263 (1886).

²When we talk about regression, we assume some law behind the data (compare with Remark 2.2).

³Stephen Senn. Francis Galton and regression to the mean. Significance, Vol. 8, Iss. 3.

⁴<https://www.randomservices.org/random/data/Galton.html>

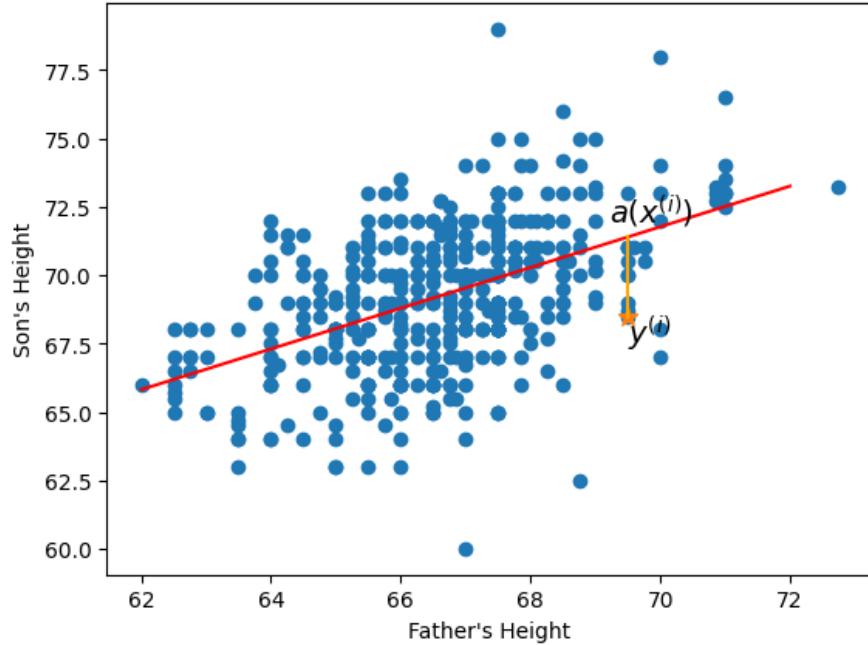


Figure 4.2: Scatter plot of the sons' height VS fathers' height from Galton's notebook

If we want to predict the children's height using the father's height using a linear function

$$a(x) = w_1 x + w_0, \quad (4.1)$$

we may want to minimize an average error or the sum of (squared) distances from this line to the real height:

$$\frac{1}{N} \sum_{i=1}^N \left(a(x^{(i)}) - y^{(i)} \right)^2 = \frac{1}{N} \sum_{i=1}^N \left(w_1 x^{(i)} + w_0 - y^{(i)} \right)^2 \rightarrow \min_{w_0, w_1}. \quad (4.2)$$

The error (4.2) depends only on two parameters w_0 and w_1 , in other words, it is a function of two variables.

4.1 Optimization of Smooth Convex Functions

Definition 4.1

A point x^* is said to be a **local minimum** of a function $f: \mathbb{R}^d \rightarrow \mathbb{R}$, if there exists a neighborhood of some radius $\varepsilon B_\varepsilon(x^*)$, such that

$$f(x^*) \leq f(x), \quad \forall x \in B_\varepsilon(x^*). \quad (4.3)$$

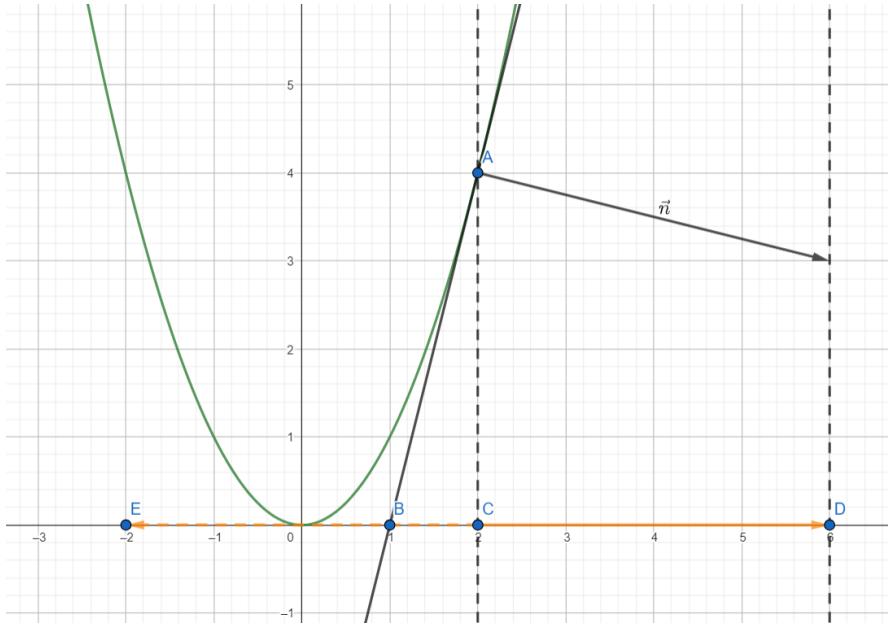


Figure 4.3: Illustration of the normal vector.

Definition 4.2

A function f is said to be **convex** if for two points x and y

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y), \quad \lambda \in [0, 1]. \quad (4.4)$$

A smooth convex function has a unique minimum.

4.2 Gradient Decent in 1D

Consider a function

$$f(x) = x^2. \quad (4.5)$$

This function is differentiable and convex. And it has a unique minimum at $x_* = 0$ (see Fig. 4.3).

Recall the equation of a line with slope k

$$y = kx + b. \quad (4.6)$$

If we know that it goes through the point (x_0, y_0) , then

$$y_0 = kx_0 + b. \quad (4.7)$$

Subtracting these equations, we get the equation of the tangent line:

$$y - y_0 = k(x - x_0) \quad \text{or} \quad \frac{x - x_0}{1} = \frac{y - y_0}{k}, \quad (4.8)$$

4 Linear Regression

where $k = f'(x_0)$. And $\vec{\ell} = (1, k)$ is the vector along this line. The unnormalized normal vector can be chosen as⁵

$$\vec{n} = (k, -1). \quad (4.9)$$

Consider a point $x_0 = 2$, the derivative at this point is equal to $f'(x_0) = 4$, and the normal vector $n = (4, -1)$ (see Fig. 4.3). The projection of the vector \vec{n} indicates the direction of the growth of the function f . This means that if we go from the point $x_0 = 2$ to the point $x_1 = x_0 + h = 6$, we will get a higher value of f .

This observation gives us the idea to go in the opposite direction, i.e., from x_0 to $x_0 - h$. In this case the new point is

$$x_1 = x_0 - f'(x_0) = 2 - 4 = -2. \quad (4.10)$$

If we continue this process and chose $x_2 = x_1 - f'(x_1) = -2 - (-4) = 2$, we will bounce between two points. To avoid this, we can multiply $f'(x_0)$ by a scaling factor $0 < \eta < 1$. This small modification of the rule (4.10) can resolve this problem. Now the update rule for every iteration takes form

$$x^{new} = x^{old} - \eta f'(x^{old}). \quad (4.11)$$

If we use $\eta = 0.5$, then $x_1 = x_0 - 0.5f'(x_0) = 2 - 0.5 \cdot 4 = 0$, and we reach the minimum of f in one iteration.

Mathematically, by definition of a differentiable function,

$$f(x + h) - f(x) = f'(x)h + o(h) \quad (4.12)$$

It means, that if we move from the point x to point $x + h$, the function will increase by $f'(x)h$ plus a term that vanishes, when $h \rightarrow 0$.

Let us set $h = -\eta f'(x)$, then

$$f(x + h) - f(x) = -\eta(f'(x))^2 + o(\eta f'(x)) \quad (4.13)$$

or

$$f(x + h) = f(x) - \eta(f'(x))^2 + o(\eta f'(x)). \quad (4.14)$$

Choosing η small enough, we can make the last term smaller than the previous one, then we have

$$f(x + h) < f(x). \quad (4.15)$$

This means that the value of the function f will decrease in each iteration and eventually it will reach the global minimum.

4.3 Optimization and Gradient Decent in 2D

4.3.1 Partial Derivatives and Gradient Vector

Consider a function

$$z = x^2 + (y - 1)^2. \quad (4.16)$$

4.3 Optimization and Gradient Decent in 2D

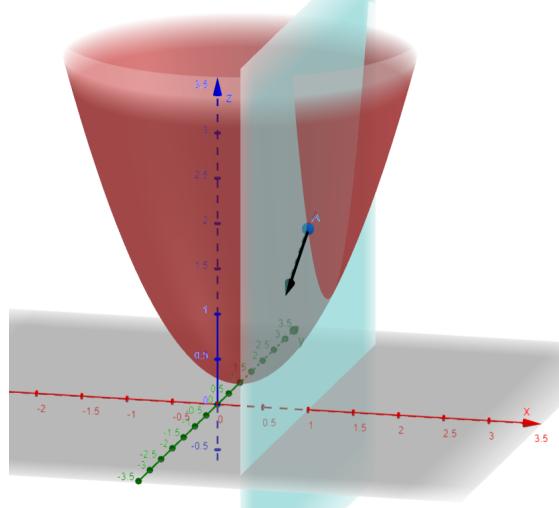


Figure 4.4: The normal vector in the section $x = 1$

Consider point $(x_0, y_0) = (1, 0)$. The value at this point is $z(1, 0) = 2$. If we fix $x = 1$, then in the section we will see a parabola:

$$z(1, y) = 1 + (y - 1)^2. \quad (4.17)$$

Because it is a function of one variable, the normal vector is

$$\vec{n}_y = \left(0, \frac{\partial z}{\partial y} \Big|_{(x,y)=(1,0)}, -1 \right). \quad (4.18)$$

When we calculate the partial derivative with respect to y , we treat x as a constant.⁶ In our case,

$$\frac{\partial z}{\partial y} = 2(y - 1) \quad \text{and} \quad \vec{n}_y = (0, -2, -1). \quad (4.19)$$

Now, we fix another coordinate: $y = 0$ (see Fig.). Then we have a function of one variable x :

$$z(x, 0) = x^2 + 1. \quad (4.20)$$

The normal vector at $x = 1$ is

$$\vec{n}_x = \left(\frac{\partial z}{\partial x} \Big|_{(x,y)=(1,0)}, 0, -1 \right). \quad (4.21)$$

⁵Because the dot product $\vec{\ell} \cdot \vec{n} = 0$.

⁶By definition, $\frac{\partial z}{\partial y} = \lim_{\Delta y \rightarrow 0} \frac{z(x, y + \Delta y) - z(x, y)}{\Delta y}$.

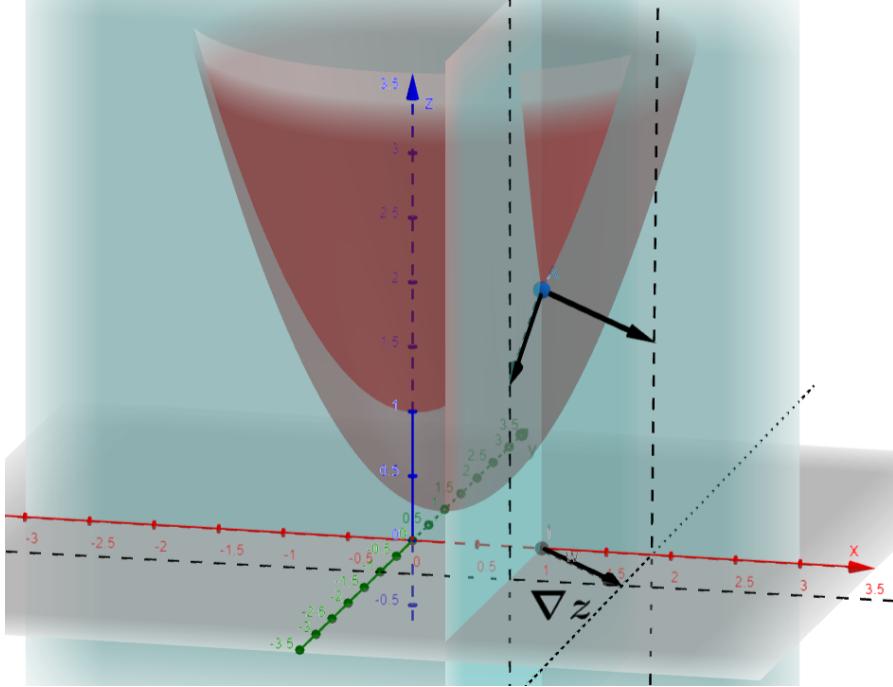


Figure 4.5: The gradient at $(x, y) = (1, 0)$ indicates the direction of the speediest increase of the function z

When we calculate the partial derivative with respect to x , we treat y as a constant. In our case,

$$\frac{\partial z}{\partial x} = 2x \quad \text{and} \quad \vec{n}_x = (0, 1, -1). \quad (4.22)$$

Definition 4.3

The vector of partial derivatives

$$\text{grad } z = \nabla z = \left(\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y} \right) \quad (4.23)$$

is called a **gradient** of the function of two variables z .

4.3.2 Minimization of a Smooth Convex Function in 2D

Because in the minimum the normal vector to the function $z = f(x, y)$

$$\vec{n} = \left(\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}, -1 \right) \quad (4.24)$$

should point down, that is equal to the zero gradient:

$$\nabla z = \left(\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y} \right) = 0. \quad (4.25)$$

In case of our function $z = x^2 + (y - 1)^2$, we have two equations:

$$\begin{cases} \frac{\partial z}{\partial x} = 2x = 0, \\ \frac{\partial z}{\partial y} = 2(y - 1) = 0. \end{cases} \quad (4.26)$$

The solution gives the point of minimum:

$$x^* = 0, y^* = 1, \quad \text{and} \quad z(x^*, y^*) = 0. \quad (4.27)$$

4.3.3 Gradient Descent Algorithm

In two-dimensional case the gradient descent algorithm looks as follows.

1. Choose the initial point (x_0, y_0)
2. Update the points by the rule

$$\begin{pmatrix} x^{new} \\ y^{new} \end{pmatrix} = \begin{pmatrix} x^{old} \\ y^{old} \end{pmatrix} - \eta \begin{pmatrix} \frac{\partial z}{\partial x} \Big|_{(x,y)=(x^{old},y^{old})} \\ \frac{\partial z}{\partial y} \Big|_{(x,y)=(x^{old},y^{old})} \end{pmatrix} \quad (4.28)$$

or in vector notation

$$\mathbf{x}^{new} = \mathbf{x}^{old} - \eta \nabla z(\mathbf{x}^{old}). \quad (4.29)$$

There are theorems about the upper bound for the learning rate η in the case of a smooth function (see, for example, Theorem 3.4 in <https://arxiv.org/abs/2301.11235>). But often we don't know the properties of the loss function, then we choose the learning rate small enough, and if the algorithm does not converge, we try smaller learning rate value.

For the loss function (4.2) the algorithm takes form

Algorithm 1 GD

```

 $T \leftarrow$  Number of epochs       $\triangleright$  The number of times we go through the entire dataset
 $w \leftarrow$  Initialize (randomly)
for epoch = 1..T do
     $w^{new} = w^{old} - \eta \nabla_w L(w^{old})$ 
    You might want to update the learning rate  $\eta$ 
end for

```

4.4 Least Mean Squares for Linear Regression

4.4.1 Simple Linear Regression

Assume, the data have only one feature. Then the model takes form

$$a(x) = w_0 + w_1 x. \quad (4.30)$$

We want to minimize the sum of squared residuals:

$$SSR(w_0, w_1) = \sum_{i=1}^N \left(y^{(i)} - w_0 - w_1 x^{(i)} \right)^2. \quad (4.31)$$

In order to find optimal values w_0^* and w_1^* , we should solve the following equations:

$$\begin{cases} \frac{\partial SSR}{\partial w_0} = -2 \sum_{i=1}^N \left(y^{(i)} - w_0 - w_1 x^{(i)} \right) = 0, \\ \frac{\partial SSR}{\partial w_1} = -2 \sum_{i=1}^N x^{(i)} \left(y^{(i)} - w_0 - w_1 x^{(i)} \right) = 0. \end{cases} \quad (4.32)$$

After division by -2 we get

$$\begin{cases} \sum_{i=1}^N y^{(i)} - Nw_0 - w_1 \sum_{i=1}^N x^{(i)} = 0, \\ \sum_{i=1}^N x^{(i)}y^{(i)} - w_0 \sum_{i=1}^N x^{(i)} - w_1 \sum_{i=1}^N x^{(i)2} = 0. \end{cases} \quad (4.33)$$

or

$$\begin{cases} \bar{y} - w_0 - w_1 \bar{x} = 0, \\ \bar{xy} - w_0 \bar{x} - w_1 \bar{x^2} = 0, \end{cases} \quad (4.34)$$

where

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x^{(i)}, \quad \bar{y} = \frac{1}{N} \sum_{i=1}^N y^{(i)}, \quad \bar{xy} = \frac{1}{N} \sum_{i=1}^N x^{(i)}y^{(i)}, \quad \bar{x^2} = \frac{1}{N} \sum_{i=1}^N x^{(i)2}. \quad (4.35)$$

After solution of this system with respect to w_0 and w_1 , we will get

$$a(x) = \bar{y} + \rho_{XY} \frac{\sigma_X}{\sigma_Y} (x - \bar{x}), \quad (4.36)$$

where

$$\rho_{XY} = \frac{\frac{1}{N} \sum_{i=1}^N (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\sigma_X \sigma_Y}, \quad \sigma_X^2 = \frac{1}{N} \sum_{i=1}^N (x^{(i)} - \bar{x})^2, \quad \sigma_Y^2 = \frac{1}{N} \sum_{i=1}^N (y^{(i)} - \bar{y})^2 \quad (4.37)$$

are correlation and variances correspondingly.

4.5 Closed Form Solution for Linear Regression

Consider the following model for data $(x^{(i)}, y^{(i)})$:

$$y^{(i)} \approx \hat{y}^{(i)} = w_0 + w_1 x_1^{(i)} + w_2 x_2^{(i)} + \dots + w_d x_d^{(i)}, \quad i = 1, 2, \dots, N. \quad (4.38)$$

Using vector notation, it can be represented as follows:

$$y^{(i)} \approx \hat{y}^{(i)} = \tilde{x}^{(i)} w^T, \quad (4.39)$$

where $\tilde{x}^{(i)} = (1, x_1^{(i)}, x_2^{(i)}, \dots, x_d^{(i)})^T$ and $w = (w_0, w_1, w_2, \dots, w_d)^T$.

The sum of squared residuals can be written using the dot product:

$$SSR = (y^{(1)} - \hat{y}^{(1)}, \dots, y^{(N)} - \hat{y}^{(N)}) \begin{pmatrix} y^{(1)} - \hat{y}^{(1)} \\ \vdots \\ y^{(N)} - \hat{y}^{(N)} \end{pmatrix}. \quad (4.40)$$

Let us introduce more convenient notation for the vector of errors:

$$\begin{aligned} \begin{pmatrix} y^{(1)} - \hat{y}^{(1)} \\ \vdots \\ y^{(N)} - \hat{y}^{(N)} \end{pmatrix} &= \begin{pmatrix} y^{(1)} \\ \vdots \\ y^{(N)} \end{pmatrix} - \begin{pmatrix} w_0 + w_1 x_1^{(1)} + w_2 x_2^{(1)} + \dots + w_d x_d^{(1)} \\ \vdots \\ w_0 + w_1 x_1^{(N)} + w_2 x_2^{(N)} + \dots + w_d x_d^{(N)} \end{pmatrix} \\ &= \begin{pmatrix} y^{(1)} \\ \vdots \\ y^{(N)} \end{pmatrix} - \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \dots & x_d^{(1)} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_1^{(N)} & x_2^{(N)} & \dots & x_d^{(N)} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{pmatrix} = y - Xw. \end{aligned} \quad (4.41)$$

Thus,

$$SSR = (y - Xw)^T (y - Xw). \quad (4.42)$$

To minimize this function, we can calculate its gradient and equate it to zero.⁷

$$\begin{aligned} \frac{\partial SSR}{\partial w} &= \frac{\partial}{\partial w} ((y^T - w^T X^T)(y - Xw)) \\ &= \frac{\partial}{\partial w} (y^T y - w^T X^T y - y^T Xw + w^T X^T Xw) \\ &= -2X^T y + 2X^T Xw = 0. \end{aligned} \quad (4.43)$$

From the last equation we find

$$w^* = (X^T X)^{-1} X^T y. \quad (4.44)$$

And the model looks as follows:

$$a(x) = w_0^* + w_1^* x_1 + \dots + w_d^* x_d = (w^*, \tilde{x}) = w^{*T} \tilde{x} = \tilde{x}^T w^*, \quad (4.45)$$

where \tilde{x} is the vector x extended by a constant one.

⁷Note that the by the derivative $\frac{\partial f}{\partial A}$ of a scalar function $f(A)$ with respect to a matrix A we mean the derivatives with respect to each component of A , i.e., the resulted shape should be the same the shape of the matrix A , it is referred as **denominator layout**. The formulas of matrix derivatives can be found at https://en.wikipedia.org/wiki/Matrix_calculus.

4.6 Regression Metrics

$$MSE = \frac{1}{N} \sum_{i=1}^N \left(a(x^{(i)}) - y^{(i)} \right)^2 \quad (4.46)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N \left(a(x^{(i)}) - y^{(i)} \right)^2} \quad (4.47)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N \left| a(x^{(i)}) - y^{(i)} \right| \quad (4.48)$$

$$Huber = \frac{1}{N} \sum_{i=1}^N \phi_\varepsilon \left(a(x^{(i)}) - y^{(i)} \right), \quad (4.49)$$

where

$$\phi_\varepsilon(z) = \begin{cases} \frac{1}{2}z^2, & |z| < \varepsilon, \\ \varepsilon \left(|z| - \frac{1}{2}\varepsilon \right), & |z| \geq \varepsilon. \end{cases}$$

$$MAPE = \frac{1}{N} \sum_{i=1}^N \left| \frac{y^{(i)} - a(x^{(i)})}{y^{(i)}} \right| \quad (4.50)$$

It allows for larger errors for large values of the target and penalizes smaller errors more heavily for small values of the target. E.g., $\frac{1000 - 900}{1000} = 0.1$ and $\frac{10 - 9}{9} = 0.1$. The first case has an absolute error of 100, while the second has an absolute error of 1. MAPE is often used in stock planning.

$$R^2 = 1 - \frac{\sum_{i=1}^N (y^{(i)} - a(x^{(i)}))^2}{\sum_{i=1}^N (y^{(i)} - \bar{y})^2} \quad (4.51)$$

If the model has a constant prediction $a(x) = \bar{y}$, $R^2 = 0$. When the prediction is 100% correct, $R^2 = 1$. This metric is returned by scikit-learn model, when you call `model.score(X, y)`.

$$Quantile = \frac{1}{N} \sum_{i=1}^N \rho_\tau \left(y^{(i)} - a(x^{(i)}) \right), \quad (4.52)$$

where

$$\rho_\tau(z) = \begin{cases} \tau z, & z > 0, \\ (\tau - 1)z, & z \leq 0. \end{cases}$$

$$MSLE = \frac{1}{N} \sum_{i=1}^N \left(\log \left(a(x^{(i)}) + 1 \right) - \log \left(y^{(i)} + 1 \right) \right)^2 \quad (4.53)$$

This metric is used when we are more interested in the relative magnitude or order of the error rather than the precise absolute error.

4.7 Statistical Linear Regression

The optimal predictor for MSE loss

$$a^* = \underset{a \in \mathcal{A}}{\operatorname{argmin}} \text{Risk}(a; \text{Data}) = \underset{a \in \mathcal{A}}{\operatorname{argmin}} E[\text{Loss}(a(X), Y)].$$

is given by

$$a^*(x) = E[Y|X = x]$$

The conditional expectation $E[Y|X = x]$ is called a **regression line**.

Under assumption that X and Y have joint Gaussian distribution

$$(X, Y) \sim N(\mu, \Sigma), \quad \mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_X & \rho \\ \rho & \sigma_Y \end{pmatrix}$$

the conditional distribution is also **Gaussian**

$$Y|X = x \sim N\left(\mu_Y + \frac{\sigma_Y}{\sigma_X}\rho(x - \mu_X), (1 - \rho^2)\sigma_Y\right)$$

Therefore, the regression line is

$$a^*(x) = E[Y|X = x] = \mu_Y + \frac{\sigma_Y}{\sigma_X}\rho(x - \mu_X) = w_0 + w_1x,$$

where

$$w_0 = \mu_Y - \frac{\sigma_Y}{\sigma_X}\rho\mu_X, \quad w_1 = \frac{\sigma_Y}{\sigma_X}\rho.$$

This means that in the case when the data $(x^{(i)}, y^{(i)})$ come from an underlying normal distribution, the ordinary least squares (OLS) estimator coincides with the statistical regression estimator.

4.8 Exercises

Problem 1 (Linear Model). Assume that we have three points:

x	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$
y	$y^{(1)}$	$y^{(2)}$	$y^{(3)}$

and we want to draw a line

$$y = w_0 + w_1x$$

minimizing the sum of squared errors

$$L(w_0, w_1) = \left(y^{(1)} - w_0 - w_1x^{(1)}\right)^2 + \left(y^{(2)} - w_0 - w_1x^{(2)}\right)^2 + \left(y^{(3)} - w_0 - w_1x^{(3)}\right)^2.$$

4 Linear Regression

This is a quadratic function; therefore, it has a unique minimum.

Calculate the derivatives:

$$\frac{\partial L}{\partial w_0} =$$

$$\frac{\partial L}{\partial w_1} =$$

Factor out w_0 and w_1 and equate the derivatives to zero:

$$\frac{\partial L}{\partial w_0} =$$

$$\frac{\partial L}{\partial w_1} =$$

Divide the equations by 6 and rewrite them using the following notation:

$$\bar{x} = \frac{x^{(1)} + x^{(2)} + x^{(3)}}{3}, \quad \bar{y} = \frac{y^{(1)} + y^{(2)} + y^{(3)}}{3},$$

$$\bar{x^2} = \frac{x^{(1)2} + x^{(2)2} + x^{(3)2}}{3}, \quad \bar{xy} = \frac{x^{(1)}y^{(1)} + x^{(2)}y^{(2)} + x^{(3)}y^{(3)}}{3}.$$

Eq.1 :

Eq.2 :

Solve the equations for w_0 and w_1

$$w_0 =$$

$$w_1 =$$

Problem 2. Given the following data

x	1	2	3
y	1	$\frac{3}{2}$	$\frac{5}{2}$

Calculate weights (result of Problem 1) and plot the line (draw the given points as well).

$$w_0 =$$

$$w_1 =$$

Problem 3. Prove formula (4.36).

Solution. From (4.34)

$$w_0 = \bar{y} - w_1 \bar{x}$$

$$w_1 = \frac{\bar{xy} - \bar{x}\bar{y}}{\bar{x}^2 - (\bar{x})^2}$$

Let's rewrite the last equality:

$$\begin{aligned}
w_1 &= \frac{\frac{1}{N} \sum_{i=1}^N x^{(i)} y^{(i)} - \left(\frac{1}{N} \sum_{i=1}^N x^{(i)}\right) \left(\frac{1}{N} \sum_{i=1}^N y^{(i)}\right)}{\frac{1}{N} \sum_{i=1}^N x^{(i)2} - \left(\frac{1}{N} \sum_{i=1}^N x^{(i)}\right)^2} = \\
&= \frac{\frac{1}{N} \sum_{i=1}^N x^{(i)} y^{(i)} - \bar{y} \frac{1}{N} \sum_{i=1}^N x^{(i)} - \bar{x} \frac{1}{N} \sum_{i=1}^N y^{(i)} + \bar{x} \frac{1}{N} \sum_{i=1}^N y^{(i)}}{\frac{1}{N} \sum_{i=1}^N x^{(i)2} - \bar{x} \frac{1}{N} \sum_{i=1}^N x^{(i)} - \bar{x} \frac{1}{N} \sum_{i=1}^N x^{(i)} + (\bar{x})^2} \\
&= \frac{\frac{1}{N} \sum_{i=1}^N (x^{(i)} y^{(i)} - \bar{y} x^{(i)} - \bar{x} y^{(i)} + \bar{x} y^{(i)})}{\frac{1}{N} \sum_{i=1}^N (x^{(i)2} - 2\bar{x} x^{(i)} + (\bar{x})^2)} \\
&= \frac{\frac{1}{N} \sum_{i=1}^N (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\left(\sqrt{\frac{1}{N} \sum_{i=1}^N (x^{(i)} - \bar{x})^2}\right)^2} \tag{4.54} \\
&= \frac{\frac{1}{N} \sum_{i=1}^N (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\left(\sqrt{\frac{1}{N} \sum_{i=1}^N (x^{(i)} - \bar{x})^2}\right)^2} \sqrt{\frac{1}{N} \sum_{i=1}^N (y^{(i)} - \bar{y})^2} \\
&= \frac{\frac{1}{N} \sum_{i=1}^N (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\sqrt{\frac{1}{N} \sum_{i=1}^N (x^{(i)} - \bar{x})^2} \sqrt{\frac{1}{N} \sum_{i=1}^N (y^{(i)} - \bar{y})^2}} \sqrt{\frac{1}{N} \sum_{i=1}^N (y^{(i)} - \bar{y})^2} = \rho_{XY} \frac{\sigma_Y}{\sigma_X}.
\end{aligned}$$

Problem 4. From expression (4.43) derive the gradient descent formula (4.29) in matrix notation:

$$w^{new} = function(X, y, w^{old}, \eta) \tag{4.55}$$

5 Stochastic Gradient Descent

5.1 SGD

The minimization of SSR is equivalent to the minimization of MSE:

$$L(w) = \frac{1}{N} SSR = \frac{1}{N} \sum_{i=1}^N \left(y^{(i)} - w^T \tilde{x}^{(i)} \right)^2 = \frac{1}{N} \sum_{i=1}^N L^{(i)}(w). \quad (5.1)$$

In the full-batch gradient descent, we update the weights by the formula

$$w^{new} = w^{old} - \eta \nabla_w L(w^{old}). \quad (5.2)$$

In stochastic gradient descent, for each update we are using only one term of the loss function, that means we use only one data point on each iteration.

Algorithm 2 SGD

```

 $T \leftarrow$  Number of epochs       $\triangleright$  The number of times we go through the entire dataset
 $w \leftarrow$  Initialize (randomly)
for epoch = 1..T do
    for n = 1..N do
         $i \leftarrow$  Sample  $i \sim U\{1, \dots, N\}$ 
         $w^{new} = w^{old} - \eta \nabla_w L^{(i)}(w^{old})$ 
    end for
    You might want to update the learning rate  $\eta$ 
end for

```

From the next statement it follows that the stochastic gradient is an unbiased estimator for the full-batch gradient.

Proposition 5.1

If we have a function represented as a sum of functions

$$f(w) = \sum_{i=1}^N f_i(w), \quad (5.3)$$

then the function $\hat{f}(\mathcal{I}; w) = N f_{\mathcal{I}}(w)$ is an unbiased estimator of $f(w)$. Here we assume that the random variable \mathcal{I} can take any of the values 1, ..., N with probability $1/N$.

5 Stochastic Gradient Descent

In the above statement, the assumption is that the only randomness is contained in the index \mathcal{I} , which has a uniform distribution on the set $\{1, 2, \dots, N\}$. Then expected value of $Nf_{\mathcal{I}}(w)$ is

$$E[Nf_{\mathcal{I}}(w)] = \sum_{i=1}^N \frac{1}{N} N f_i(w) = f(w). \quad (5.4)$$

To make it more precise, consider the data $(x^{(1)}, y^{(1)}), \dots, (x^{(N)}, y^{(N)})$ and function

$$L(w) = \sum_{i=1}^N L^{(i)}(w) = \sum_{i=1}^N \left(y^{(i)} - w^T x^{(i)} \right)^2. \quad (5.5)$$

Consider a random variable \mathcal{I} uniformly distributed on $\{1, 2, \dots, N\}$ and the estimator

$$\hat{L}(\mathcal{I}; w) = NL^{(\mathcal{I})} = N \left(y^{(\mathcal{I})} - w^T x^{(\mathcal{I})} \right)^2 \quad (5.6)$$

with corresponding estimate

$$\hat{L}(\hat{i}; w) = N \left(y^{(\hat{i})} - w^T x^{(\hat{i})} \right)^2, \quad \hat{i} \sim U\{1, 2, \dots, N\}. \quad (5.7)$$

Then its expected value is equal to

$$E[\hat{L}(\mathcal{I}; w)] = \sum_{i=1}^N \frac{1}{N} N \left(y^{(i)} - w^T x^{(i)} \right)^2 = L(w). \quad (5.8)$$

5.2 Mini-Batch SGD

In the gradient descent algorithm, we use all the data points for one step, in the stochastic gradient descent, we use one point at a time. There is a compromise solution: we can use some amount M of data points for one step. This approach is known as *mini batch SGD*.

$$Loss_{SGD} = \left(a(x^i) - y^{(i)} \right)^2$$

$$Loss_{mini-batch} = \frac{1}{M} \sum_{i=1}^M \left(a(x^i) - y^{(i)} \right)^2$$

If we minimize

$$Loss_{batch} = \frac{1}{N} \sum_{i=1}^N \left(a(x^i) - y^{(i)} \right)^2,$$

it is called full-batch SGD or simply gradient descent (GD).

The theorems about convergence of SGD and Minibatch SGD can be found in Sec. 5 and 6 in <https://arxiv.org/abs/2301.11235>.

Algorithm 3 Minibatch SGD

```

 $T \leftarrow$  Number of epochs       $\triangleright$  The number of times we go through the entire dataset
 $M \leftarrow$  Batch size
 $w \leftarrow$  Initialize (randomly)
for epoch = 1..T do
    You may shuffle the data
    for batch = 1..N/M do
         $L_{batch} \leftarrow \frac{1}{M} \sum_{i=(batch-1)M}^{batch \cdot M} L^{(i)}$ 
         $w^{new} \leftarrow w^{old} - \eta \nabla_w L_{batch}(w^{old})$ 
    end for
    You might want to update the learning rate  $\eta$ 
end for

```

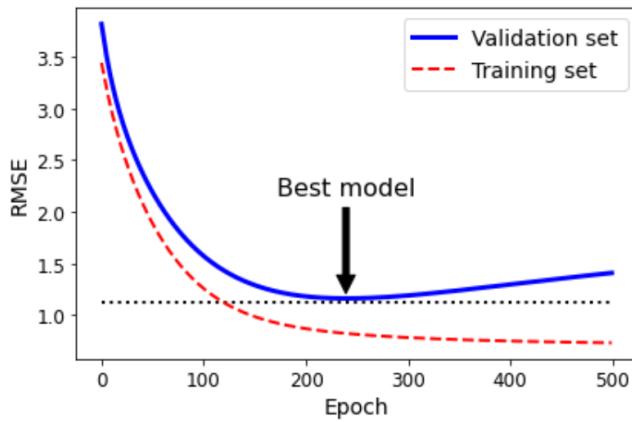


Figure 5.1: Typical plots for training and validation loss

5.3 Early Stopping

In SGD algorithm, the iterations go on for the number of epochs. But we can check the performance of our model on the validation set. Typically, it will be improving, but at some point it will grow up again (see Fig. 5.1).¹ This is the time to stop training the model, because it will fit the training data more accurately, but lose the performance on the unseen data. This is called *memorization* of the data or *overfitting*.

5.4 Regularization

The idea of regularization appeared in the problem of inversion of matrices. For example, a matrix A with all zero elements is not invertible, but $A + \varepsilon I$ will be invertible if ε is greater than the machine's epsilon.

¹https://github.com/ageron/handson-ml2/blob/master/04_training_linear_models.ipynb

5 Stochastic Gradient Descent

Consider the following model for the apartments price prediction:

$$a(x) = 10 + 100 \cdot \text{number_rooms}. \quad (5.9)$$

It can work correctly for one bedroom apartments, but it will give an unreasonable increment if the number of bedrooms is changed.

We can add a so called *regularization term* to our loss function, which will force the weights to be smaller.

In **ridged regression** we add ℓ_2 norm of the weights:

$$\text{Loss} + \alpha \|w_{-0}\|_2^2 \rightarrow \min$$

In **Lasso regression**² ℓ_1 norm of the weights is used:

$$\text{Loss} + \beta \|w_{-0}\|_1 \rightarrow \min$$

Elastic regression combines both ℓ_2 and ℓ_1 terms:

$$\text{Loss} + \alpha \|w_{-0}\|_2^2 + \beta \|w_{-0}\|_1 \rightarrow \min$$

Usually, the bias term w_0 is not included, i.e., $\|w_{-0}\|_1 = |w_1| + \dots + |w_d|$ and $\|w_{-0}\|_2^2 = w_1^2 + \dots + w_d^2$. Otherwise, the regression line (or plain) would go through the origin.

See class notebook for the choice of the regularization and parameters α and β .

5.5 Optimization for Ridge Regression

We minimize the following function:³

$$\text{Loss}(w) = \text{SSR}(w) + \alpha \|w_{-0}\|^2, \quad (5.10)$$

where we use the following notation:

$$\|w_{-0}\|^2 = w_1^2 + \dots + w_d^2. \quad (5.11)$$

We write the loss function in matrix notation:

$$\begin{aligned} \text{Loss}(w) &= \|y - Xw\|^2 + \alpha \|w_{-0}\|^2 \\ &= (y - Xw)^T (y - Xw) + \alpha (w_1^2 + \dots + w_d^2) \\ &= y^T y - 2y^T Xw + w^T X^T Xw + \alpha w^T I_{-0} w \\ &= y^T y - 2y^T Xw + w^T (X^T X + \alpha I_{-0}) w, \end{aligned} \quad (5.12)$$

²From Least Absolute Shrinkage and Selection Operator. See the next chapter for the explanation of this name.

³The factor $1/N$ is typically omitted at the first term.

where

$$I_{-0} = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix}. \quad (5.13)$$

$$\frac{\partial Loss}{\partial w} = -2X^T y + 2(X^T X + \alpha I_{-1})w = 0. \quad (5.14)$$

Thus,

$$w = (X^T X + \alpha I_{-0})^{-1} X^T y. \quad (5.15)$$

5.6 Optimization for Lasso Regression

We minimize the following function:

$$Loss(w) = SSR(w) + \alpha \|w_{-0}\|_1, \quad (5.16)$$

where we use the following notation:

$$\|w_{-0}\|_1 = |w_1| + \dots + |w_d|. \quad (5.17)$$

This function does not have a gradient on the line $w_1 = \dots = w_d = 0$. One possible solution is to use the proximal gradient descent method.

5.6.1 Proximal Gradient Descent

Proposition 5.2

In the gradient descent method, when we minimize a function $f(x)$, the update step

$$y = x_t - \eta \nabla f(x_t) \quad (5.18)$$

is equivalent to the following optimization problem:

$$f(x_t) + (\nabla f(x_t))^T (y - x_t) + \frac{1}{2\eta} \|y - x_t\|^2 \rightarrow \min_y. \quad (5.19)$$

The first term in (5.19) doesn't depend on y ; therefore, it's equivalent to

$$\frac{\eta}{2} \|\nabla f(x_t)\|^2 + (\nabla f(x_t))^T (y - x_t) + \frac{1}{2\eta} \|y - x_t\|^2 \rightarrow \min_y. \quad (5.20)$$

And after multiplication by 2η

$$\|\eta \nabla f(x_t)\|^2 + 2(\nabla f(x_t))^T (y - x_t) + \|y - x_t\|^2 = \|y - (x_t - \eta \nabla f(x_t))\|^2 \rightarrow \min_y. \quad (5.21)$$

If we want to minimize the sum

$$f(x) + g(x) \quad (5.22)$$

5 Stochastic Gradient Descent

for the next approximation, we can solve the following problem:

$$\frac{1}{2\eta} \|y - (x_t - \eta \nabla f(x_t))\|^2 + g(y) \rightarrow \min_y. \quad (5.23)$$

The first term minimizes $f(x)$ and the second term minimizes $g(x)$.

Denote by

$$\text{prox}_{g,\eta}(z) = \operatorname{argmin}_y \left[\frac{1}{2\eta} \|y - z\|^2 + g(y) \right]. \quad (5.24)$$

Then the update step can be written as

$$x_{t+1} = \text{prox}_{g,\eta}(x_t - \eta \nabla f(x_t)). \quad (5.25)$$

In case of Lasso regression,

$$f(w) = \frac{1}{2} \text{SSR}, \quad g(w) = \|w_0\|_1. \quad (5.26)$$

In this case,

$$\text{prox}_{g,\eta}(z) = \operatorname{sign}(z) \max(|z| - \alpha\eta, 0). \quad (5.27)$$

You can watch the following videos for the calculation of the proximal gradient <https://www.youtube.com/watch?v=LABiFM79W6M> and implementation Lasso regression <https://www.youtube.com/watch?v=0Vt1PJKAjxg>

A more rigorous solution and understanding of the *shrinkage* can be get by introducing a new variable z and reducing the minimization of (5.16) to the following minimization problem with constrains:

$$F(w, z) = \|z\|^2 + \alpha \|w_0\|_1 \rightarrow \min_{w,z}, \quad (5.28)$$

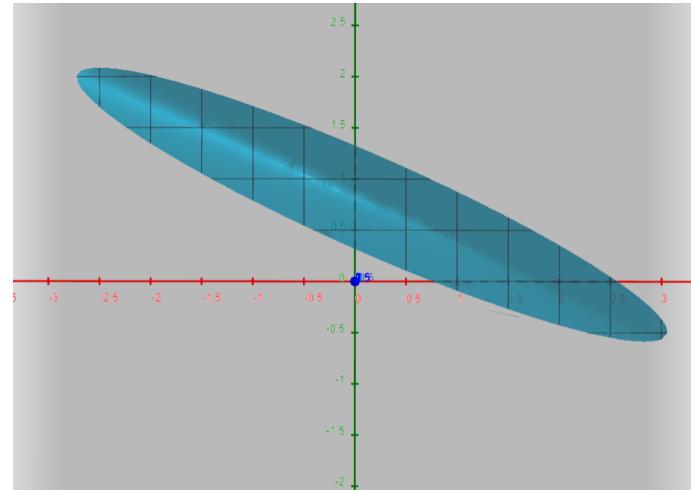
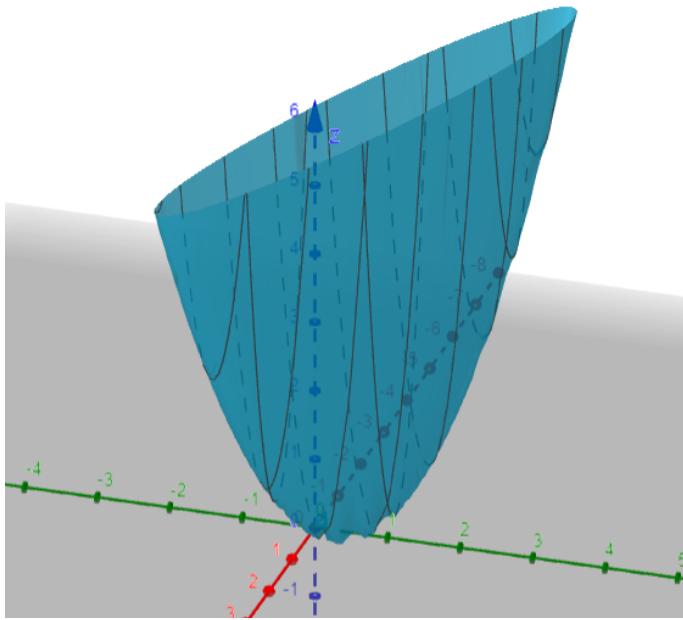
$$y - Xw = z. \quad (5.29)$$

And apply Karush — Kuhn — Tucker Theorem (see Ch. 7).

5.7 Exercises

Problem 3. The plots of the SSR for Problem 2 is equal to

$$L(w_0, w_1) = L_1(w_0, w_1) + L_2(w_0, w_1) + L_3(w_0, w_1).$$



Problem 3. Apply one epoch of SGD. Initialize $w_0^{old} = w_1^{old} = 0$.

Calculate the derivatives of $L_1(w_0, w_1) = (y^{(1)} - w_0 - w_1 x^{(1)})^2$ (at the first point):

$$\frac{\partial L_1}{\partial w_0} =$$

$$\frac{\partial L_1}{\partial w_1} =$$

Update the weights (use the learning rate $\eta = 1$):

$$w_0^{new} = w_0^{old} - \eta \frac{\partial L_1}{\partial w_0} =$$

$$w_1^{new} =$$

Calculate the derivatives of $L_2(w_0, w_1) = (y^{(2)} - w_0 - w_1 x^{(2)})^2$:

$$\frac{\partial L_2}{\partial w_0} =$$

$$\frac{\partial L_2}{\partial w_1} =$$

Update the weights:

$$w_0^{new} =$$

$$w_1^{new} =$$

Calculate the derivatives of $L_3(w_0, w_1) = (y^{(3)} - w_0 - w_1 x^{(3)})^2$:

$$\frac{\partial L_3}{\partial w_0} =$$

$$\frac{\partial L_3}{\partial w_1} =$$

Update the weights:

$$w_0^{new} =$$

$$w_1^{new} =$$

5 Stochastic Gradient Descent

Plot each point in the figure on the right.

Problem 4. Can you think of how to improve convergence?

6 Model Selection

By *model selection* we usually mean the choice of the model including hyper-parameters. For example, in case of regression we can use kNN regression and tune the number of neighbors k , we can try the ridged regression with loss function

$$L(w) = \frac{1}{N} \sum_{i=1}^N \left(y^{(i)} - w^T \tilde{x}^{(i)} \right)^2 + \alpha \|w\|_2^2. \quad (6.1)$$

Then we should tune the parameter α .

This can be done by *grid search*: we train the model for different values of parameters and compare the error on the validation data set or by cross-validation. We create a set of values for parameter $\alpha : \alpha_1, \alpha_2, \dots, \alpha_{S_1}$, a set of values for another parameter: $\beta_1, \beta_2, \dots, \beta_{S_2}$ and so on. All possible combinations of those values create a grid.

Sometimes, the choice of features is also considered as part of the model selection.

6.1 Feature Selection

6.1.1 By Criterion

We can use correlation

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}} \quad (6.2)$$

or mutual information (MI)¹

$$I(X, Y) = D_{\text{KL}}(p(x, y) || p(x)p(y)) \quad (6.3)$$

of each feature with the target and take the features with the criterion value higher than a predefined threshold.

You can try different thresholds by using a model and evaluating the error on the validation set.

Mutual information works better because correlation indicates linear dependence only.

¹The Kullback — Leibler divergence measure the discrepancy between two distributions. Therefore, mutual information shows the strength of dependence of two random variables. It is zero, when they are independent, i.e., when $p(x, y) = p(x)p(y)$.

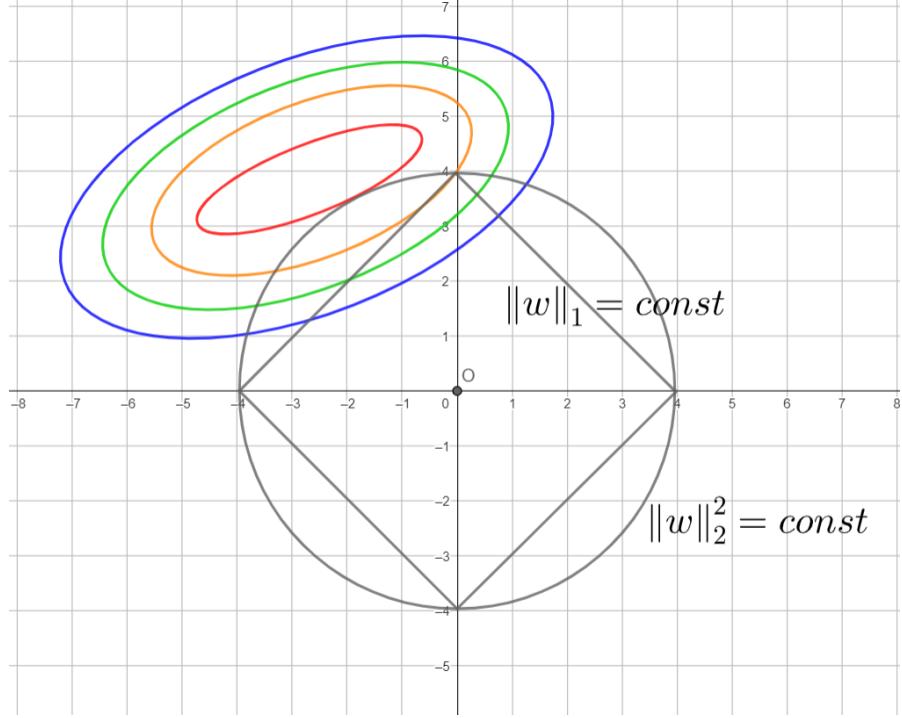


Figure 6.1: Counter plots of the loss function and ℓ_1 and ℓ_2 regularization terms with the same value

6.1.2 By Model

Lasso regularization is sometimes referred to as a *sparse* regression:

$$L(w) = \text{MSE}(w) + \alpha \|w\|_1. \quad (6.4)$$

If the position of the MSE loss function is not symmetric around the origin (see Fig. 6.1), then for a fixed value of the regularization term

$$\alpha \|w\|_1 = C_2, \quad (6.5)$$

its minimum may lie in the corner of the ℓ_1 ball (which has a constant value C_2 on its boundary). Assume that the loss function has some value C :

$$L(w) = \text{MSE}(w) + \alpha \|w\|_1 = C_1 + C_2 = C. \quad (6.6)$$

If we consider the ℓ_2 regularization, because the ℓ_2 ball is symmetric in any direction, typically, the minimum will be reached at some point with non-zero w_1 and w_2 components. In case of the ℓ_1 ball, the minimum will be reached with $w_1 = 0$. This means that the feature with index 1 is not important and can be omitted.

We can train Lasso regression and choose the features with the highest absolute values of the weights.

6.1.3 By (Recursive) Elimination

In this approach, we try all possible combinations of r features. The number r can be chosen by the model error on the validation set.

7 Linear Models for Binary Classification

Assume that we are given points $(x^{(i)}, y^{(i)})$, $i = 1, 2, \dots, N$ (see Fig. 7.1), and $y^{(i)} \in \{-1, 1\}$. In 3D, those points will lie in the plane $z = -1$ for one class and on the planes $z = \pm 1$ (see Fig. 7.2).

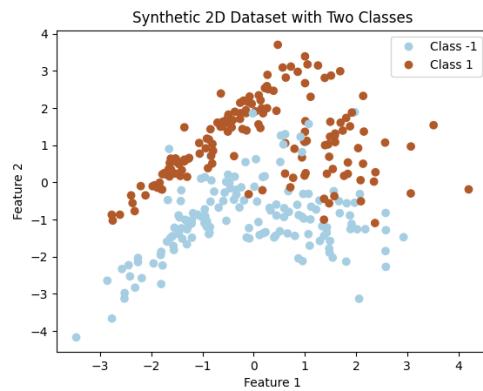


Figure 7.1: Caption

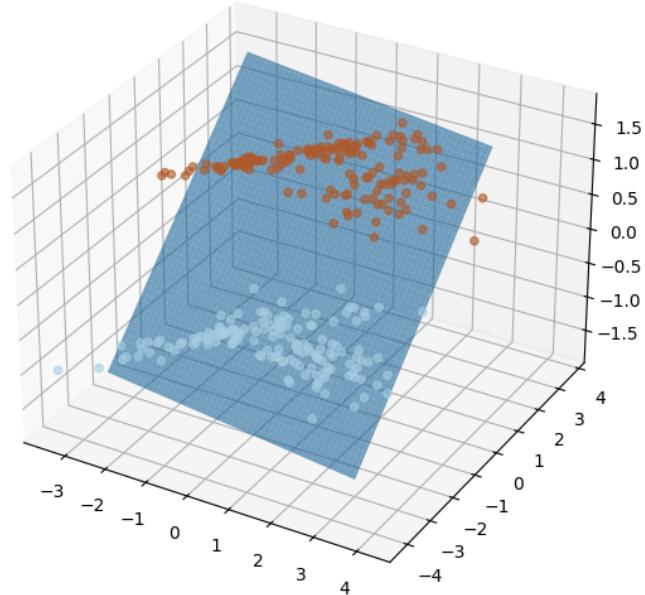


Figure 7.2: Caption

We have a linear model

$$a(x) = w_0 + w_1 x_1 + w_2 x_2.$$

which returns values from $(-\infty, \infty)$. We can modify it to return -1 or 1 :

$$a(x) = \text{sign}(w_0 + w_1 x_1 + w_2 x_2) = \text{sign } w^T \tilde{x} = \text{sign} \langle w, \tilde{x} \rangle, \quad \tilde{x} = (1, x_1, x_2). \quad (7.1)$$

7.1 Geometry of Linear Classifier

Equation

$$w^T \tilde{x} = 0 \quad (7.2)$$

defines a line in the plain $z = w^T \tilde{x}$. If we look at this figure from above, the projection of this line on the feature space will look like a separator. Every point below this line is classified as class -1 and every point above is classified as class $+1$.

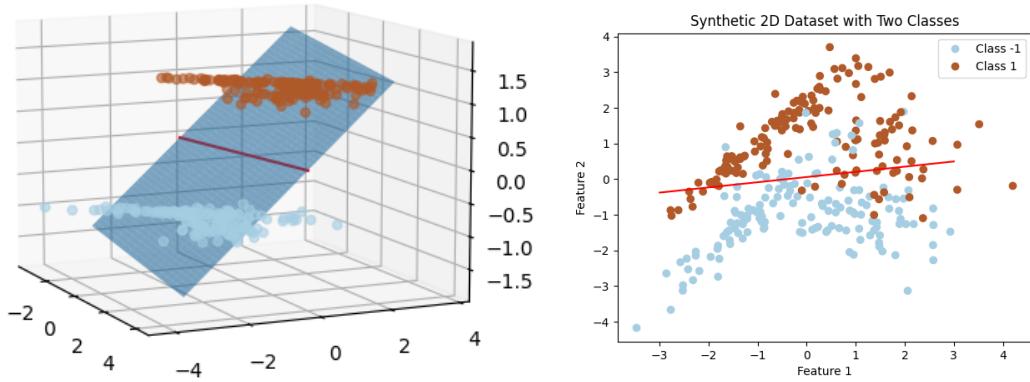


Figure 7.3: Threshold is set to 0

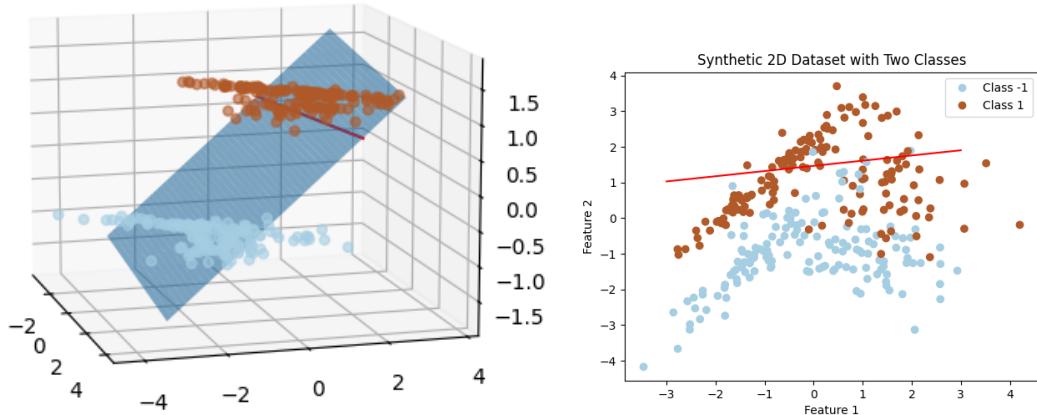


Figure 7.4: Threshold is set to 0.7

We can consider a different equation:

$$w^T \tilde{x} = th, \quad (7.3)$$

that corresponds to a change of the bias term $w'_0 = w_0 - th$:

$$w_0 - th + w_1 x_1 + w_2 x_2 = 0. \quad (7.4)$$

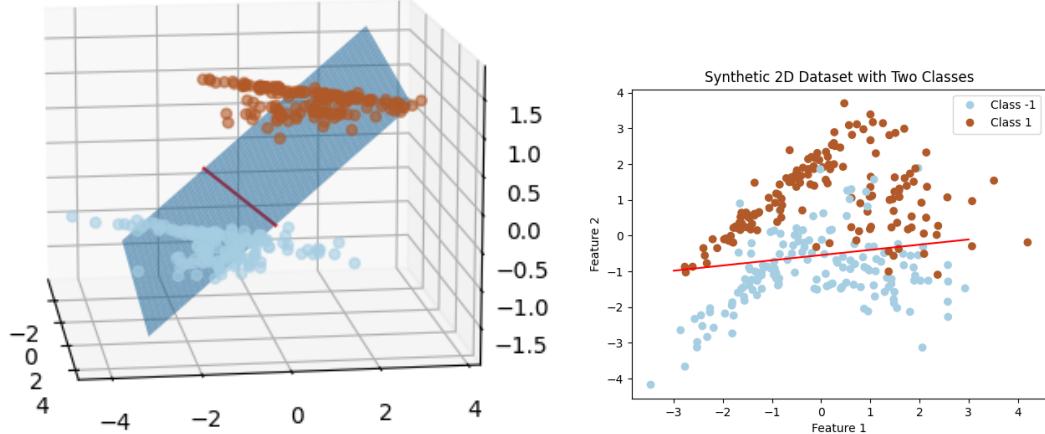
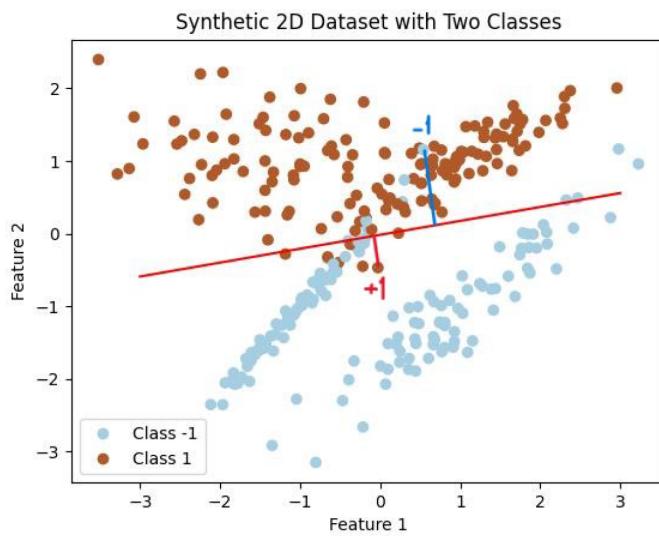


Figure 7.5: Threshold is set to -0.3

7.2 Margin

The distance from a point (x_1, x_2) above the line to the line is given by (see Problem 4)

$$d = \frac{w_0 + w_1 x_1 + w_2 x_2}{\sqrt{w_1^2 + w_2^2}} = \frac{\langle w, \tilde{x} \rangle}{\sqrt{w_1^2 + w_2^2}},$$



By **margin** of $x^{(i)}$ we mean the distance multiplied by the sign $y^{(i)}$:

$$y \frac{\langle w, \tilde{x} \rangle}{\sqrt{w_1^2 + w_2^2}}, \quad (7.5)$$

If $y = -1$ and $\langle w, \tilde{x} \rangle > 0$, then $y\langle w, \tilde{x} \rangle < 0$

If $y = -1$ and $\langle w, \tilde{x} \rangle < 0$, then $y\langle w, \tilde{x} \rangle > 0$

If $y = 1$ and $\langle w, \tilde{x} \rangle < 0$, then $y\langle w, \tilde{x} \rangle < 0$

If $y = 1$ and $\langle w, \tilde{x} \rangle > 0$, then $y\langle w, \tilde{x} \rangle > 0$

For the correct classification the margin is positive. That is why sometimes we call by **margin** the value

$$y\langle w, \tilde{x} \rangle, \quad (7.6)$$

which is positive for correct predictions and negative otherwise.

7.3 Support Vector Machine (SVM)

In this section, we use a slightly different notation: $w = (w_1, w_2)$ instead of $w = (w_0, w_1, w_2)$.

7.3.1 Separable Case

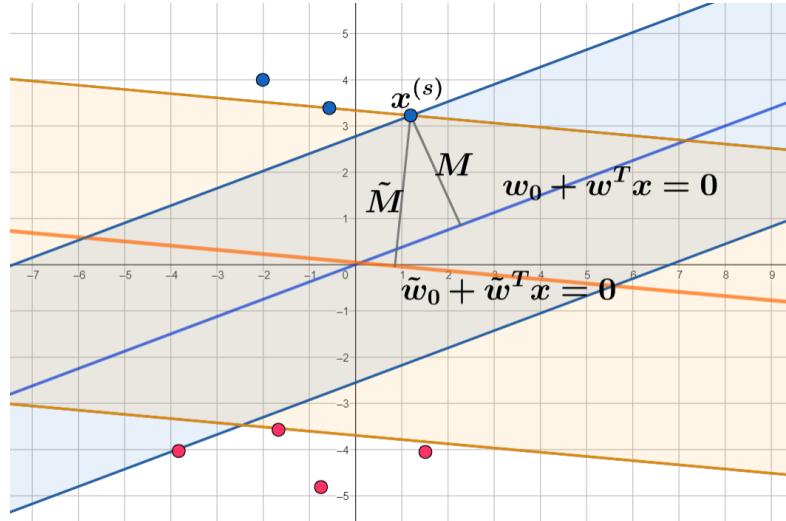


Figure 7.6: Two different separators

In Fig. 7.6, there are two separating lines:

$$w_0 + w^T x = 0 \quad \text{and} \quad \tilde{w}_0 + \tilde{w}^T x = 0. \quad (7.7)$$

7.3 Support Vector Machine (SVM)

If we move the separating line up, at some point we will reach a point from our data and we won't be able to move the line further. The points that do not allow to move the separating line are called **supporting** points (vectors).

For a separable case the margin should be positive:

$$\frac{y^{(i)}(w_0 + w^T x^{(i)})}{\|w\|} > 0, \quad i = 1, \dots, N. \quad (7.8)$$

We assume that the best separator has the maximal possible margin to for the closest point from the decision line:

$$M(w_0, w) = \min_{i=1, \dots, N} \frac{y^{(i)}(w_0 + w^T x^{(i)})}{\|w\|} = \frac{1}{\|w\|} \min_{i=1, \dots, N} y^{(i)}(w_0 + w^T x^{(i)}) \rightarrow \max_{w_0, w}. \quad (7.9)$$

In addition, we assume the normalization constrain: for the support boundary (the line containing supporting points) equation looks as¹

$$w_0 + w^T x = 1. \quad (7.10)$$

Under this assumption problem (7.9) takes form

$$M(w_0, w) = \frac{1}{\|w\|} \rightarrow \max. \quad (7.11)$$

From normalization condition (7.11):

$$\min_{i=1, \dots, N} y^{(i)}(w_0 + w^T x^{(i)}) = 1. \quad (7.12)$$

Thus, we have the constrained optimization problem

$$\begin{cases} M(w_0, w) \rightarrow \max, \\ y^{(i)}(w_0 + w^T x^{(i)}) \geq 1, \quad i = 1, \dots, N. \end{cases} \quad (7.13)$$

or, equivalently,

$$\begin{cases} \frac{1}{2} \|w\|^2 \rightarrow \min, \\ y^{(i)}(w_0 + w^T x^{(i)}) \geq 1, \quad i = 1, \dots, N. \end{cases} \quad (7.14)$$

¹For fixed separator $w_0 + w^T x = 0$ this scaling can be done by division by $w_0 + w^T x^{(s)}$: $\tilde{w}_0 + \tilde{w}^T x = 0$, with $\tilde{w}_0 = \frac{1}{w_0 + w^T x^{(s)}}$ and $\tilde{w} = \frac{1}{w_0 + w^T x^{(s)}}$ describes the same line, but $\tilde{w}_0 + \tilde{w}^T x^{(s)} = 1$.

7.3.2 Non-Separable Case

We can allow for some points to lie inside the margin (see Fig. 7.7) by introducing non-negative *slack* variables:

$$\begin{cases} \frac{1}{2} \|w\|^2 \rightarrow \min_{w_0, w}, \\ y^{(i)}(w_0 + w^T x^{(i)}) \geq 1 - \xi^{(i)}, \quad i = 1, \dots, N, \\ \xi^{(i)} \geq 0, \quad i = 1, \dots, N. \end{cases} \quad (7.15)$$

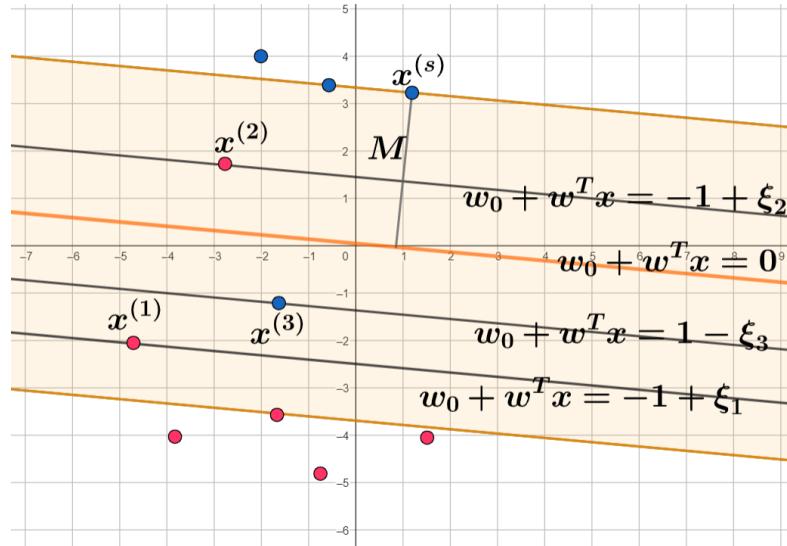


Figure 7.7: Separation with slack variables

But we can run into a problem like in Fig. 7.8, where decision boundary goes through the lowest point of one class and allows large negative margins for the other class. In order to avoid this, the additional term in the minimizing function is introduced:

$$\begin{cases} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi^{(i)} \rightarrow \min_{w_0, w, \xi^{(i)}}, \\ y^{(i)}(w_0 + w^T x^{(i)}) \geq 1 - \xi^{(i)}, \quad i = 1, \dots, N, \\ \xi^{(i)} \geq 0, \quad i = 1, \dots, N. \end{cases} \quad (7.16)$$

It forces the slack variables to be as small controlling by the constant $C > 0$.

To solve this problem KKT theorem will be used.

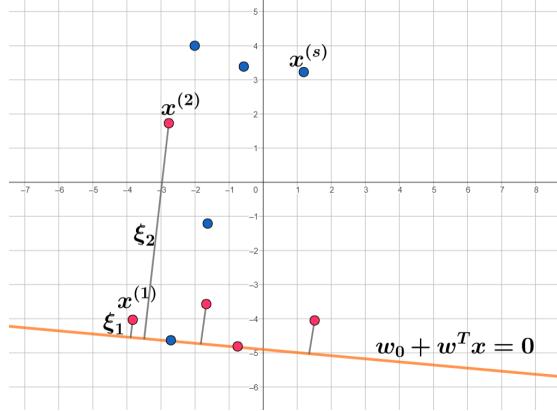


Figure 7.8: Separation without minimizing the slack variables

7.3.3 Karush — Kuhn — Tucker Theorem

Primal Problem

$$\begin{cases} f(x) \rightarrow \min_x, \\ g_i(x) \leq 0, \\ h_j(x) = 0. \end{cases} \quad (7.17)$$

The Dual Problem

$$\begin{cases} \inf_x \mathcal{L}(x, \lambda, \mu) \rightarrow \max_{\lambda, \mu \geq 0}, \\ \sum_i \lambda_i g_i(x) = 0. \end{cases} \quad (7.18)$$

where \mathcal{L} is called a Lagrangian:

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \sum_i \lambda_i g_i(x) + \sum_j \mu_j h_j(x). \quad (7.19)$$

If the functions f , g_i , and h_j are convex and there exists a point $x^{interior}$ satisfying the conditions given by inequalities and equalities, then the saddle point of the dual problem gives the solution to the primal problem and the slackness condition is satisfied:

$$\sum_i \lambda_i g_i(x) = 0.$$

We apply the Karush — Kuhn — Tucker theorem to problem (7.16):

$$\mathcal{L}(w_0, w, \xi, \lambda, \mu) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi^{(i)} - \sum_{i=1}^N \lambda_i \left(y^{(i)} (w_0 + w^T x^{(i)}) - 1 + \xi^{(i)} \right) - \sum_{i=1}^N \mu_i \xi^{(i)} \quad (7.20)$$

Calculate the derivatives:

$$\frac{\partial \mathcal{L}}{\partial w_0} = - \sum_{i=1}^N \lambda_i y^{(i)} = 0, \quad (7.21)$$

$$\frac{\partial \mathcal{L}}{\partial w} = w - \sum_{i=1}^N \lambda_i y^{(i)} x^{(i)} = 0, \quad (7.22)$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = C - \lambda_i - \mu_i = 0 \Rightarrow \lambda_i + \mu_i = C. \quad (7.23)$$

And slackness conditions

$$\lambda_i \left(y^{(i)} (w_0 + w^T x^{(i)}) - 1 + \xi_i \right) = 0, \quad (7.24)$$

$$\mu_i \xi_i = 0. \quad (7.25)$$

1. $\lambda_i = 0 \xrightarrow{(7.23)} \mu_i = C \xrightarrow{(7.25)} \xi^{(i)} = 0 \Rightarrow x^{(i)}$ doesn't influence separation, because not included in (7.22).
2. $0 < \lambda_i < C \xrightarrow{(7.23)} \mu_i \neq 0 \Rightarrow \xi^{(i)} = 0.$ (Or $\lambda_i = C$ and $\xi^{(i)} = 0.$) Therefore,

$$\left(y^{(i)} (w_0 + w^T x^{(i)}) - 1 \right) = 0$$

and $x^{(i)}$ is a support vector.

3. $\lambda_i = C$ and $\xi^{(i)} > 0$, then $x^{(i)}$ is on the other side of the support boundary $w_0 + w^T x^{(i)} = \pm 1.$

From equality (7.22) it follows that problem (7.16) can be reduced to the following quadratic optimization problem in a cube with side C :

$$\begin{cases} \frac{1}{2} \sum_{i,j=1}^N \lambda_i \lambda_j y^{(i)} y^{(j)} \langle x^{(i)}, x^{(j)} \rangle \rightarrow \max_{\lambda_i}, \\ 0 \leq \lambda_i \leq C, \quad i = 1, \dots, N, \\ \sum_{i=1}^N \lambda_i y^{(i)} = 0. \end{cases} \quad (7.26)$$

There are efficient solvers for such problems. To determine w_0 one can use equation (7.24).

For smaller values of C the model can make larger errors allowing bigger $\xi^{(i)}$ (see Fig. 7.9-7.11)

7.3.4 Minimization with a Hinge Loss

The other way to train SVM is to use SGD with so called *hinge loss*

$$L(M) = \max(0, 1 - M) \quad (7.27)$$

7.3 Support Vector Machine (SVM)

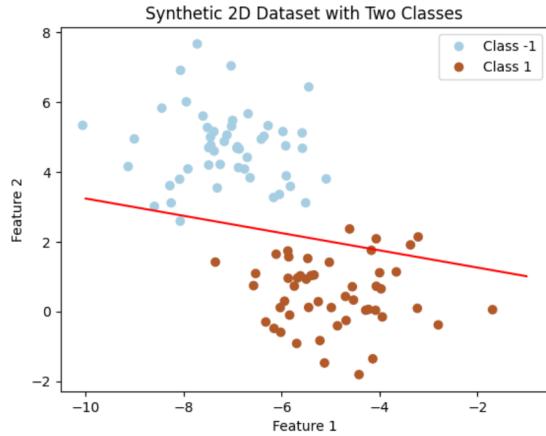


Figure 7.9: $C = 0.1$

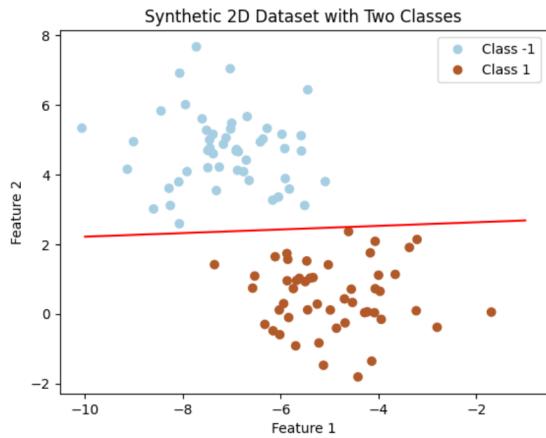


Figure 7.10: $C = 1$

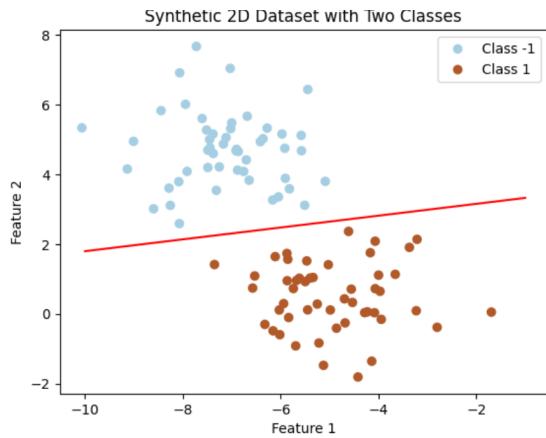


Figure 7.11: $C = 5$

7 Linear Models for Binary Classification

with ℓ_2 regularization:

$$C \sum_{i=1}^N \max \left\{ 0, 1 - y^{(i)} \left(w_0 + \langle w, x^{(i)} \rangle \right) \right\} + \frac{1}{2} \|w\|^2 \rightarrow \min_{w_0, w}. \quad (7.28)$$

It follows from the second constrain in (7.16) that

$$\xi^{(i)} \geq 1 - y^{(i)} \left(w_0 + \langle w, x^{(i)} \rangle \right) \quad (7.29)$$

And because $\xi^{(i)} \geq 0$,

$$\xi^{(i)} \geq \max \left\{ 0, 1 - y^{(i)} \left(w_0 + \langle w, x^{(i)} \rangle \right) \right\}. \quad (7.30)$$

The result of minimization of the hinge loss (7.28) is different from the solution of problem (7.26), but as practice shows, this difference is small.

7.4 F. Rosenblatt's Perceptron (1962)



C. Bishop, Pattern Recognition and Machine Learning, Sec. 4.1.7

<https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf>

7.4.1 Perceptron Algorithm

$$Loss = \frac{1}{N} \sum_{i=1}^N \left[\text{sign } a(x^{(i)}) \neq y^{(i)} \right] = \frac{1}{N} \sum_{i=1}^N \left[y^{(i)} \langle w, x^{(i)} \rangle < 0 \right]$$

Perceptron:

$$Loss = \sum_{i: y^{(i)} \langle w, x^{(i)} \rangle < 0} y^{(i)} \langle w, x^{(i)} \rangle \rightarrow \max$$

or

$$Q(w) = - \sum_{i: y^{(i)} \langle w, x^{(i)} \rangle < 0} y^{(i)} \langle w, x^{(i)} \rangle \rightarrow \min.$$

$$Q(w) = - \sum_{i: y^{(i)} \langle w, x^{(i)} \rangle < 0} y^{(i)} \langle w, x^{(i)} \rangle \rightarrow \min.$$

SGD implies the update, if i th point is missclassified:

$$w^{new} = w^{old} + y^{(i)} x^{(i)}$$

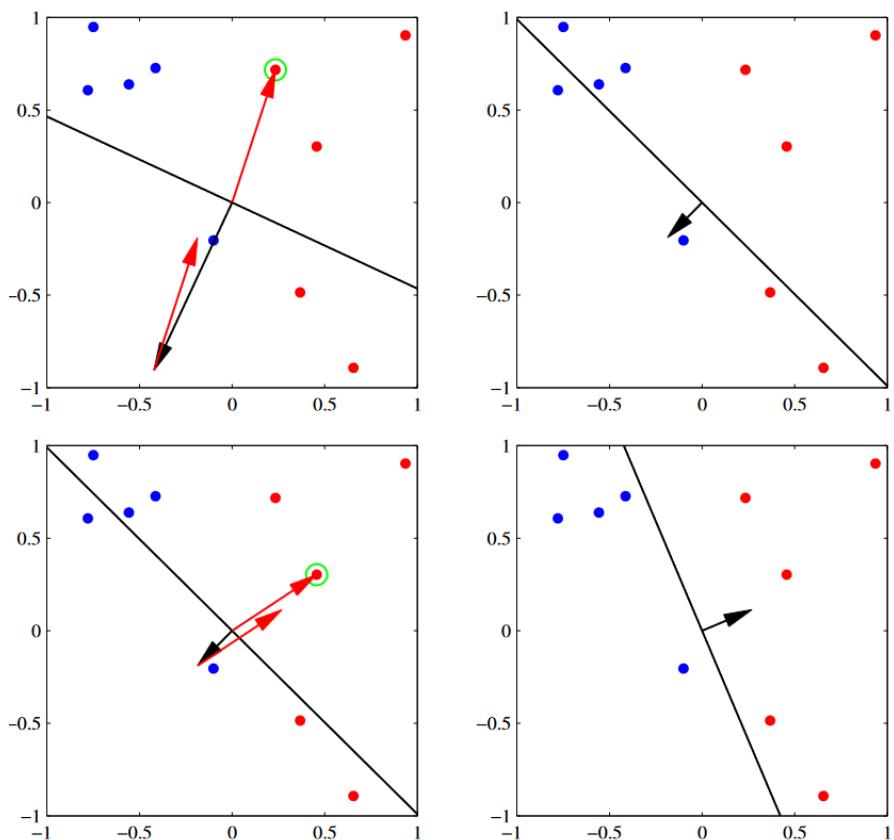


Figure 7.12: Perceptron updates

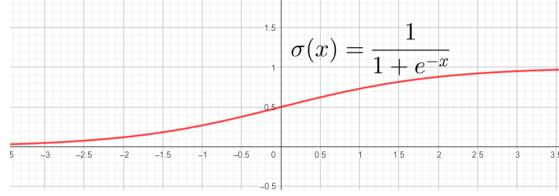


Figure 7.13: Sigmoid function

7.5 Logistic Regression

7.5.1 Sigmoid Function

In previous sections, we introduced the notion of the distance with a sign:

$$d(x) = \frac{w_0 + w^T x}{\|w\|}. \quad (7.31)$$

Let's consider a function that maps $(-\infty, +\infty)$ onto the interval $(0, 1)$:

$$\sigma: (-\infty, +\infty) \rightarrow (0, 1) \quad (7.32)$$

$$\sigma(d) = \frac{e^d}{1 + e^d} = \frac{1}{1 + e^{-d}} \quad (7.33)$$

Then we can interpret this value as probability of the positive class:

- For large values of $d > 0$ $\sigma(d) \approx 1$.
- For $d = 0$, $\sigma(d) = 0.5$.
- For large negative d , $\sigma(d) \approx 0$.

But for now, this transform doesn't have any probabilistic model behind it.

We can solve equation (7.33) with respect to d :

$$d = \log \frac{\sigma}{1 - \sigma}. \quad (7.34)$$

Because σ represents the probability of the positive class, this value is the log ratio of the probabilities of two classes. It is known as **log-odds** or **logit**.

7.5.2 Probabilistic Model

We are given the data $(x^i, y^{(i)})$, where $y^{(i)} \in \{-1, +1\}$. For every x consider a random variable Y that takes two values: $+1$ with probability $p(x)$ and -1 with probability $1 - p(x)$. And assume

Y	-1	1
$P(Y)$	$1 - p$	p

Table 7.1: Variant of Bernoulli random variable

that this probability is given by²

$$p(x) = \sigma(w_0 + w_1 x_1 + \dots + w_d x_d). \quad (7.35)$$

If the margin is positive and $y = 1$,

$$P(Y = 1; x, w) = \sigma(\langle w, \tilde{x} \rangle) = \frac{1}{1 + e^{-\langle w, \tilde{x} \rangle}} = \frac{1}{1 + e^{-y \cdot \langle w, \tilde{x} \rangle}}.$$

If the margin is positive, but $y = -1$, we have

$$P(Y = -1; x, w) = 1 - \sigma(\langle w, \tilde{x} \rangle) = 1 - \frac{1}{1 + e^{-\langle w, \tilde{x} \rangle}} = \frac{1}{1 + e^{\langle w, \tilde{x} \rangle}} = \frac{1}{1 + e^{y \cdot (-\langle w, \tilde{x} \rangle)}}.$$

The model takes the form as in Table 7.2. The features x are assumed to be known, the

Y	-1	1
$P(Y)$	$\sigma(-\langle w, \tilde{x} \rangle)$	$\sigma(\langle w, \tilde{x} \rangle)$

Table 7.2: Logistic regression model

weights w are parameters that should be tuned.

7.5.3 Maximum Likelihood

The Likelihood function is defined as follows:

$$L(y^{(1)}, y^{(2)}, \dots, y^{(N)}; \mathbf{X}, w) = \prod_{i=1}^N P(Y^{(i)} = y^{(i)}; x^{(i)}, w) \rightarrow \max_w \quad (7.36)$$

Using the representation of probabilities from Table 7.2, the likelihood takes form

$$L(w) = \prod_{i=1}^N \sigma(y^{(i)} \langle w, x^{(i)} \rangle) = \prod_{i=1}^N \frac{1}{1 + e^{-y^{(i)} \langle w, x^{(i)} \rangle}} \rightarrow \max_w. \quad (7.37)$$

We can maximize its logarithm instead:

$$\log L(w) = \sum_{i=1}^N \log \sigma(y^{(i)} \langle w, x^{(i)} \rangle) = - \sum_{i=1}^N \log(1 + e^{-y^{(i)} \langle w, x^{(i)} \rangle}) \rightarrow \max_w, \quad (7.38)$$

or minimize the negative log-loss:

$$-\log L(w) = \sum_{i=1}^N \log(1 + e^{-y^{(i)} \langle w, x^{(i)} \rangle}) \rightarrow \min_w. \quad (7.39)$$

²This is a particular case of generalized linear models (GLM), when we assume that the mean value is represented by a function of a linear function. If the labels were 0 and 1 (instead of -1 and 1), then p represents the expected value of the Bernoulli random variable.

Definition 7.1

The **binary cross-entropy loss**

$$Loss = \frac{1}{N} \sum_{i=1}^N \log \left(1 + e^{-y^{(i)} \langle w, x^{(i)} \rangle} \right) \quad (7.40)$$

is typically used in classification tasks.

The definition of entropy, cross-entropy, and KL-divergence are as follows

$$H(p) = - \sum_{i=1}^K p_i \log p_i, \quad H(p, q) = - \sum_{i=1}^K p_i \log q_i, \quad D_{KL}(p, q) = \sum_{i=1}^K p_i \log \frac{q_i}{p_i}. \quad (7.41)$$

If follows that

$$H(p, q) = H(p) + D_{KL}(p, q). \quad (7.42)$$

Therefore, the minimization of the negative cross-entropy is equivalent to the minimization of KL-divergence (we assume that the distribution p is given).

7.5.4 Other Metrics for Logistic Regression: Deviance, AIC, BIC

$$D(w) = -2(\log L(w) - \log L(w_s)) = -2 \log L(w)$$

$$AIC = -\frac{2}{N} \log L + \frac{2d}{N}$$

$$BIC = -2 \log L + d \log N$$

7.6 Classification Metrics

$$\sum_{i=1}^N [a(x^{(i)} \neq y^{(i)})] \rightarrow min$$

$$Accuracy = \sum_{i=1}^N [a(x^{(i)} = y^{(i)})] \rightarrow max$$

This metric depends on the balance of classes. For example, consider $N = 100$ data points with 95 points labeled $+1$ and 5 points labeled -1 . If our model predicts every point as positive class, then the accuracy will be 95%.

Therefore, if q is the fraction of the largest class, we want

$$Accuracy \in [q, 1].$$

Another way to access the model's performance is the error matrix.

		Predicted condition	
		Positive (PP)	Negative (PN)
		Total population = P + N	
Actual condition	Positive (P)	True positive (TP), hit	False negative (FN), type II error, miss, underestimation
	Negative (N)	False positive (FP), type I error, false alarm, overestimation	True negative (TN), correct rejection

Figure 7.14: The matrix of errors

7.6.1 Confusion Matrix

From this matrix we can calculate accuracy as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \quad (7.43)$$

This matrix contains false positive (FP) and false negative (FN) errors.

Example (Medical exam). If the test is positive and the person is sick, we can look at the following fraction

$$\text{precision} = \frac{TP}{TP + FP}.$$

This ratio says how accurate our model is when it predicts a positive class. If precision is 100%, it means that model predicts positive class without errors. If precision is 10%, it means that from 100 positive predictions, in 90 cases the person didn't have the disease.

The other type of errors arises when the model gives negative result, while the person is sick. To measure these errors, another metric is used:

$$\text{recall/sensitivity/TPR} = \frac{TP}{TP + FN} = \frac{TP}{\text{Total positive}} = \frac{TP}{P}.$$

It shows how well the model covers the positive class.



Figure 7.15: Large Precision VS Large Recall

7.6.2 *F*-score

If we want to combine the information about precision and recall in one metric, we can use so-called F_β -score.

$$F_\beta = (1 + \beta) \frac{Precision \times Recall}{\beta(Precision + Recall)}, \quad \beta \in (0, \infty). \quad (7.44)$$

Small β optimises more Precision and larger β emphasizes Recall. Recall is considered β times more important than Precision (see Problem 7). Typically, $\beta = 1$ is used:

$$F_1 = 2 \frac{Precision \times Recall}{(Precision + Recall)}. \quad (7.45)$$

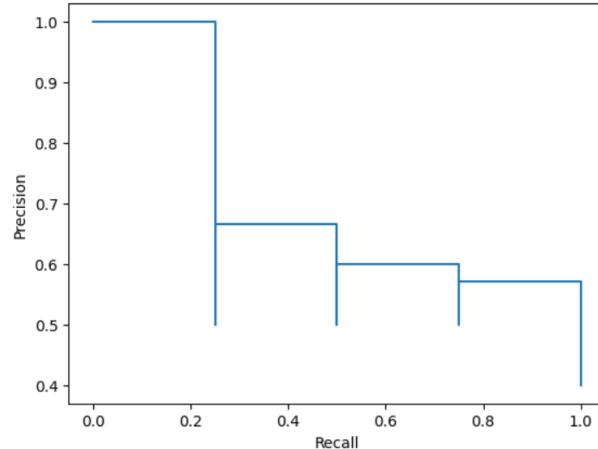
7.7 Model Selection for Classification

We can regulate precision or recall choosing the threshold $a(x) > th$. **The model can be evaluated independently of the choice of the threshold.**

7.7.1 Precision — Recall (PR) Curve

a(x)	-500	-200	-10	1	100	300	700	800	900	1000
y	-1	-1	1	-1	-1	-1	1	1	-1	1

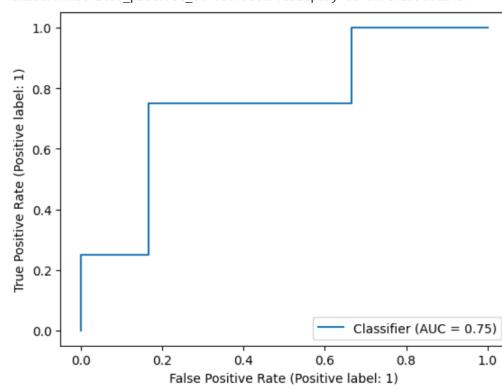
We can plot



- If the model after sorting with respect to distance often changes the classes, it has a small area under PR curve.
- if after sorting one class follows the other class, then PR curve goes through the point $(1, 1)$.

7.7.2 Receiver Operating Characteristic (ROC) Curve

```
from sklearn.metrics import roc_curve, RocCurveDisplay
RocCurveDisplay.from_predictions([-1, -1, 1, -1, -1, -1, 1, 1, -1, 1], list(range(10)))
<sklearn.metrics._plot.roc_curve.RocCurveDisplay at 0x7c7295a0e140>
```



$$Y: TPR = \frac{TP}{TP + FN} = \frac{TP}{P} \quad X: FPR = \frac{FP}{FP + TN} = \frac{FP}{N}$$

7.7.3 ROC-AUC

- Scalable to the classes size
- AUC-ROC = 1 for the ideal model
- AUC-ROC = 0.5 for the worst model
- Interpretation: fraction of the all possible ordered pairs (number of inversions the model performed).

x	$a_1(x)$	y	\hat{y}
$x^{(1)}$	10	1	1
$x^{(2)}$	9	-1	1
$x^{(3)}$	8	1	1
$x^{(4)}$	7	1	1
$x^{(5)}$	6	-1	-1
$x^{(6)}$	5	-1	-1
$x^{(7)}$	4	-1	-1
$x^{(8)}$	3	1	-1
$x^{(9)}$	2	-1	-1
$x^{(10)}$	1	-1	-1

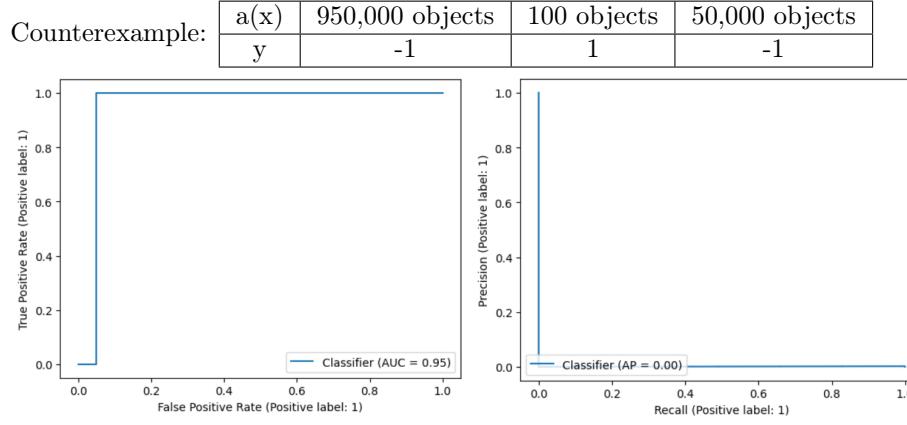
x	$a_2(x)$	y	\hat{y}
$x^{(1)}$	10	1	1
$x^{(3)}$	9	1	1
$x^{(4)}$	8	1	1
$x^{(8)}$	7	1	1
$x^{(5)}$	6	-1	-1
$x^{(6)}$	5	-1	-1
$x^{(7)}$	4	-1	-1
$x^{(2)}$	3	-1	-1
$x^{(9)}$	2	-1	-1
$x^{(10)}$	1	-1	-1

If we consider the total number of possible positive-negative pairs and the number of inversions (how many times the negative label was ranked higher than the positive one), then

$$\frac{6}{4 \times 6} = 0.25.$$

And ROC-AUC represents the ratio of the number of times the positive label was ranked higher over the total number of pairs.

7 Linear Models for Binary Classification



$$\text{AUC-ROC} = 0.95$$

$$\text{AUC-PRC} = 0.001$$

7.7.4 Quality Improvement Reporting

old r_1	0.8	0.5	0.001
new r_2	0.9	0.75	0.01
Absolute $r_2 - r_1$	0.1	0.25	0.009
Relative $\frac{r_2 - r_1}{r_1}$	12%	50%	900%

7.8 Exercises

Problem 1. Consider the line through the origin:

$$w_1x_1 + w_2x_2 = 0.$$

Prove that the vector $w = (w_1, w_2)^T$ is perpendicular to this line.

Problem 2. Prove that a distance from a point (x_1, x_2) above the line to the line is given by

$$d = \frac{\langle w, x \rangle}{|w|} = \frac{w_1x_1 + w_2x_2}{\sqrt{w_1^2 + w_2^2}}.$$

Hint: by definition, $\langle \vec{a}, \vec{b} \rangle = |\vec{a}| |\vec{b}| \cos \alpha$. Therefore, $\Pr_{\vec{a}} \vec{b} = |\vec{b}| \cos \alpha = \frac{\langle \vec{a}, \vec{b} \rangle}{|\vec{a}|}$.

Problem 3. The following value is called **margin**

$$y^{(i)} \langle w, x^{(i)} \rangle.$$

Prove that for correctly classified point $x^{(i)}$ the margin is positive and for an incorrectly classified point it's negative.

Problem 4. Assuming the separating line has a bias term

$$w_0 + w_1 x_1 + w_2 x_2 = 0,$$

prove that a distance from a point (x_1, x_2) above the line to the line is given by

$$d = \frac{w_0 + \langle w, x \rangle}{|w|} = \frac{\langle \tilde{w}, \tilde{x} \rangle}{|\tilde{w}|} = \frac{w_0 + w_1 x_1 + w_2 x_2}{\sqrt{w_1^2 + w_2^2}},$$

where $w = (w_1, w_2)^T$, $\tilde{w} = (w_0, w_1, w_2)^T$, $x = (x_1, x_2)^T$, and $\tilde{x} = (1, x_1, x_2)^T$

Problem 5. Given the following predictions and threshold $th = 700$ draw the confusion matrix and calculate precision and recall

$a(x) = \langle w, x \rangle$	y	\hat{y}
1000	1	1
900	-1	1
800	1	1
700	1	-1
300	-1	-1
100	-1	-1
1	-1	-1
-10	1	-1
-200	-1	-1
-500	-1	-1

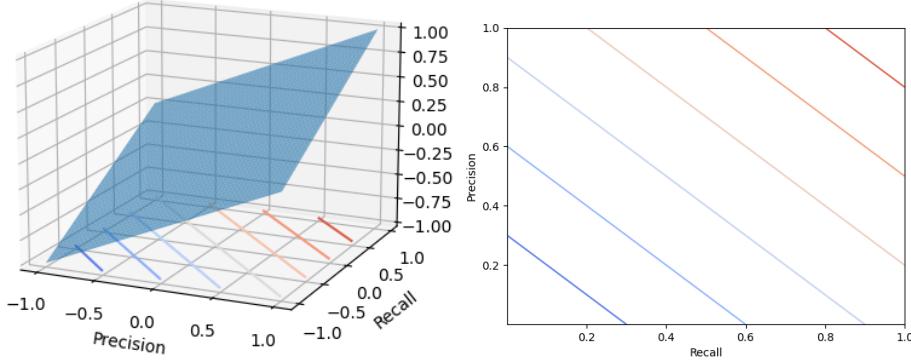
Problem 6. Given the following predictions and threshold $th = -200$ draw the confusion matrix and calculate precision and recall

$a(x) = \langle w, x \rangle$	y	\hat{y}
1000	1	1
900	-1	1
800	1	1
700	1	1
300	-1	1
100	-1	1
1	-1	1
-10	1	1
-200	-1	-1
-500	-1	-1

Problem 7. We can get better precision or recall by choosing the threshold. What if we want to combine them and maximize them together? Consider mean and harmonic mean. The surfaces and level curves are below. Why using mean is not a good idea?

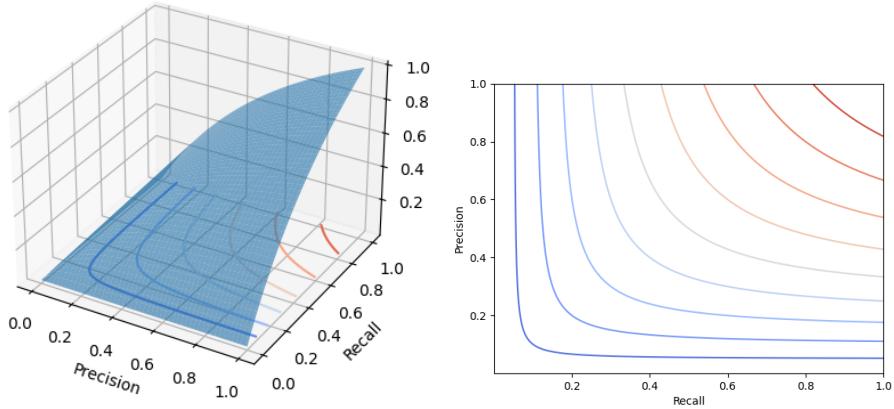
$$\frac{1}{2}(Precision + Recall)$$

7 Linear Models for Binary Classification



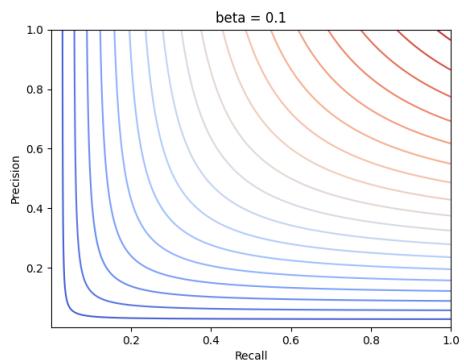
F -score

$$\frac{2(Precision \times Recall)}{Precision + Recall}$$



F_β -score. Small β optimises more Precision and larger β emphasises Recall. Recall is considered β times more important than Precision.

$$(1 + \beta) \frac{Precision \times Recall}{\beta(Precision + Recall)}$$

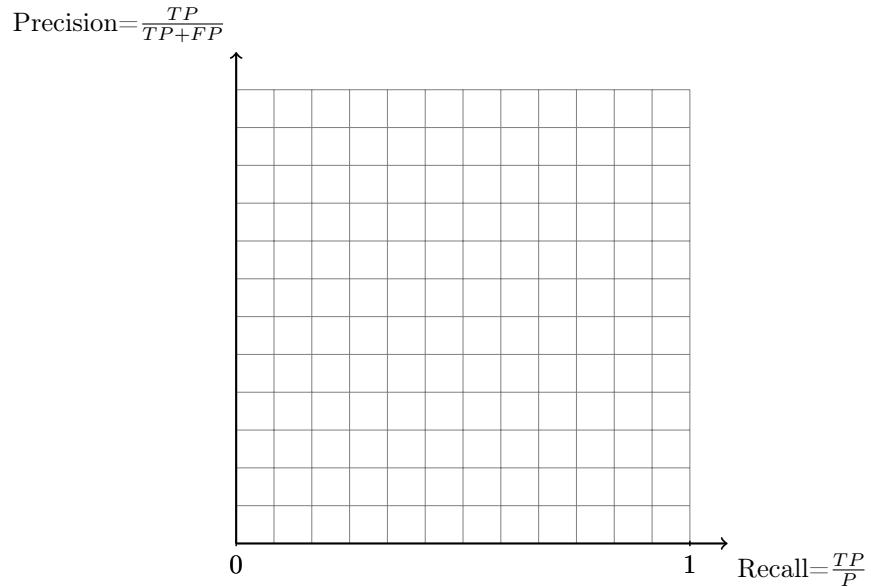


Problem 8. Calculate F_1 -score for Problem 5 and Problem 6.

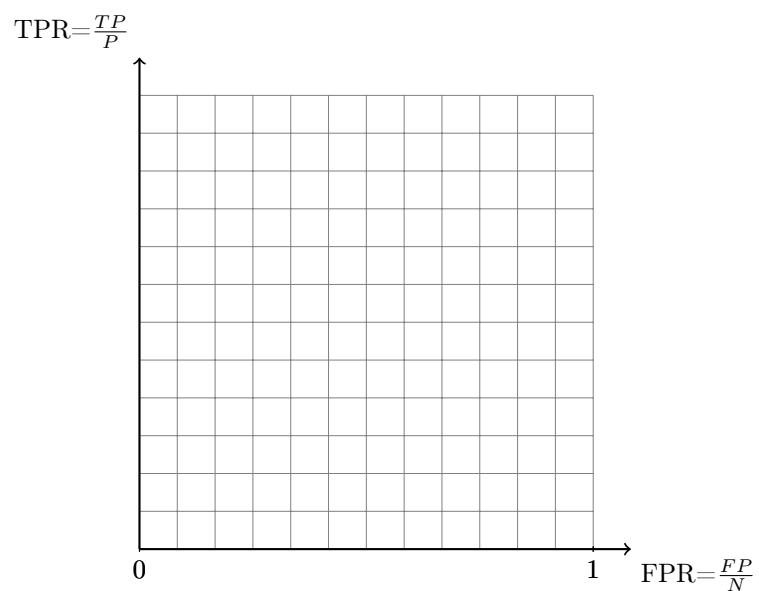
Problem 9. Given the following data

$\langle w, x \rangle$	0.45	-0.1	2	0.3	-0.5	0.7	0
y	+1	-1	+1	+1	-1	-1	+1

Plot the PR curve.



Problem 10. Using the data from Problem 9 plot ROC curve and calculate AUC.



$$\text{AUC ROC} =$$

8 Multiclass Classification

8.1 One-VS-All and All-VS-All

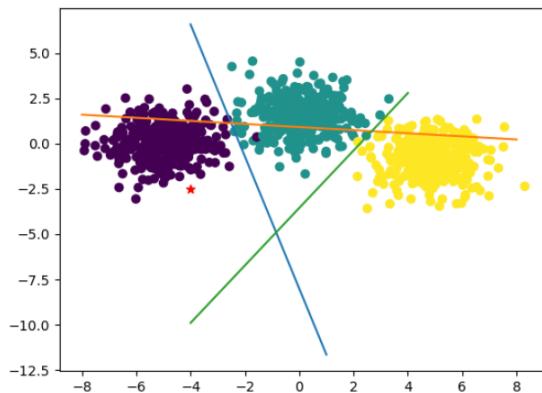
8.1.1 One - VS - all

We have

$$(X^{(i)}, y^{(i)}) \text{, where } y \in \mathbb{Y}$$
$$\mathbb{Y} = \{1, 2, \dots, C\}$$

- Train C models a_c on $(X^{(i)}, [y^{(i)} = c])$
- a_c returns distance $\langle w, x \rangle$ or probability
- The final model is

$$a(x) = \operatorname{argmax}_{c=1,2,\dots,C} a_c(x).$$



```
clf1.predict_proba(np.array([[-4, -2.5]]))[:,1], clf2.predict_proba(np.array([[-4, -2.5]]))[:,1]
```

Figure 8.1: One-VS-all classification

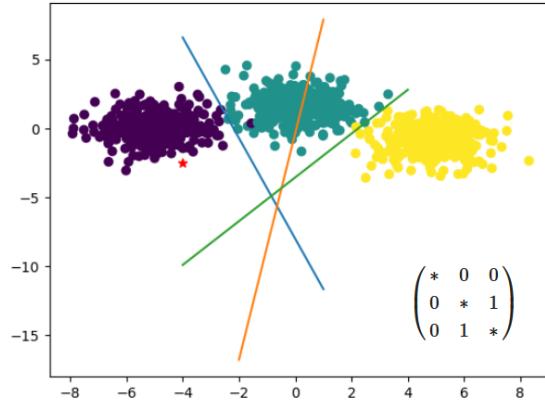


Figure 8.2: All-VS-all classification

8.1.2 All - VS - all

- For every c_1 and c_2 take $(X^{(i)}, y^{(i)})$, where $y \in \{c_1, c_2\}$
- Train model a_{c_1, c_2} on the set above
- a_{c_1, c_2} returns class c_1 or c_2
- The final model is

$$a(x) = \underset{c_1=1,2,\dots,C}{\operatorname{argmax}} \sum_{c_2=1}^C [a_{c_1, c_2}(x) = c_1].$$

8.2 Micro- and Macro-Averaging

Quality Metrics

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^N [a(x^{(i)}) = y^{(i)}]$$

Micro Averaging	Macro Averaging
TP_c, TN_c, FP_c, FN_c	
$TP = \sum_{c=1}^C TP_c, \dots$	$Precision_c = \frac{TP_c}{TP_c + FP_c}$
$Precision = \frac{TP}{TP + FP}$	$Precision = \frac{1}{C} \sum_{c=1}^C Precision_c$
Depends on the largest class	Ignores the class size

8.3 Multiple Logistic Regression

Multinomial Logistic Regression. Softmax Consider C linear models:

$$\begin{aligned} a_1(x) &= w_{1,0} + w_{1,1}x_1 + w_{1,2}x_2 + \dots + w_{1,d}x_d \\ a_2(x) &= w_{2,0} + w_{2,1}x_1 + w_{2,2}x_2 + \dots + w_{2,d}x_d \\ \dots & \\ a_C(x) &= w_{C,0} + w_{C,1}x_1 + w_{C,2}x_2 + \dots + w_{C,d}x_d \end{aligned}$$

Convert the predictions into a vector of probabilities with components

$$p_{w_c}(x) = \frac{e^{a_c(x)}}{\sum_{k=1}^C e^{a_k(x)}}$$

The likelihood looks as follows:

$$L(y^{(1)}|x^{(1)}, \dots, y^{(N)}|x^{(N)}; w_1, \dots, w_C) = \prod_{i=1}^N p_{w_1}(x^{(i)})^{[y^{(i)}=1]} \cdot \dots \cdot p_{w_C}(x^{(i)})^{[y^{(i)}=C]}$$

$$\log L = \sum_1^N \left([y^{(i)} = 1] \log p_{w_1}(x^{(i)}) + [y^{(i)} = 2] \log p_{w_2}(x^{(i)}) + \dots + [y^{(i)} = C] \log p_{w_C}(x^{(i)}) \right).$$

Because $[y^{(i)} = c] = 1$ only when $c = y^{(i)}$,

$$\log L = \sum_{i=1}^N \sum_{c=1}^C [y^{(i)} = c] \log p_{w_c}(x^{(i)}) = \sum_{i=1}^N \log p_{w_{y^{(i)}}}(x^{(i)}) \rightarrow \max_{w_1, \dots, w_C}.$$

8.3.1 Cross-Entropy Loss

If we multiply the log-likelihood by $-1/N$, then we get so called *cross entropy loss*:

$$Loss = -\frac{1}{N} \log L = -\sum_{i=1}^N \left([y^{(i)} = 1] \log p_{w_1}(x^{(i)}) + [y^{(i)} = 2] \log p_{w_2}(x^{(i)}) + \dots + [y^{(i)} = C] \log p_{w_C}(x^{(i)}) \right).$$

This name comes from the notion of cross entropy between two discrete distributions $\{p_j\}_{j=1}^J$ and $\{q_j\}_{j=1}^J$:

$$H(p, q) = -E_p[\log q] = -\sum_{j=1}^J p_j \log q_j,$$

because the loss can be represented as follows:

$$Loss = -\sum_{i=1}^N \frac{1}{N} E_{p^{(i)}}[\log q^{(i)}] = -E_i \left[E_{p^{(i)}}[\log q^{(i)}] \right] = H(Y|\mathbf{X}, Z|\mathbf{X}) \rightarrow \min_{w_1, \dots, w_C},$$

where $p^{(i)}$ and $q^{(i)}$ are discrete distributions

$$\begin{array}{c|c|c|c|c|c} Y^{(i)}|x^{(i)} & 1 & 2 & \dots & y^{(i)} & \dots & C \\ \hline p^{(i)} & 0 & 0 & \dots & 1 & \dots & 0 \end{array} \quad \begin{array}{c|c|c|c|c} Z^{(i)}|x^{(i)} & 1 & 2 & \dots & C \\ \hline q^{(i)} & p_{w_1} & p_{w_2} & \dots & p_{w_C} \end{array}$$

Recall that the smallest value of the cross entropy is obtained when $p_j = q_j$. In other words, we want to learn parameters w_c which make the distribution Z as close as possible to the empirical distribution Y .

8.4 Bayesian Classification

Suppose that we have data (x, y) , where $y \in \{1, \dots, C\}$. And x and y are samples from random variables X and Y .

By the conditional probability formula,

$$p(Y = 1|X = x) = \frac{p(X = x, Y = 1)}{p(x)} = \frac{p(X = x|Y = 1)p(Y = 1)}{p(x)}. \quad (8.1)$$

If we rewrite it as

$$p(Y = 1|X = x) = \frac{p(X = x|Y = 1)p(Y = 1)}{p(x|Y = 1)p(Y = 1) + p(x|Y = 2)p(Y = 2) + \dots + p(x|Y = C)p(Y = C)}, \quad (8.2)$$

then we get the Bayes formula. The probability $p(Y)$ is called *prior*, the probability $p(Y|X)$ is called *posterior*.

The decision rule we will use is called *maximum a posteriori* estimator (MAP) or *maximum posterior* (MP):

$$a(x) = \operatorname{argmax}_c p(Y = c|X = x). \quad (8.3)$$

Because the denominator is the same for all posterior probabilities, we can take maximum of the numerators only:

$$a(x) = \operatorname{argmax}_c p(X = x|Y = c)p(Y = c). \quad (8.4)$$

This is called also *Poor Man's Bayes*.

8.4.1 Naïve Bayes

In some cases, it is hard to estimate the distribution of multivariate features $p(X|Y)$. Then the assumption of independence is made:

$$p(X_1, \dots, X_d|Y) = p(X_1|Y) \cdot \dots \cdot p(X_d|Y). \quad (8.5)$$

For example, in spam classification problem, we can calculate the frequencies of words from vocabulary V in both classes: $p(w_i|ham)$ and $p(w_i|spam)$ ($i = 1, \dots, |V|$). Then the probability of an e-mail is

$$p(text|ham) = \prod_{w \in text} p(w|ham)^{N(w, text)}, \quad (8.6)$$

where $N(w, text)$ is the count of the word w in the text.

The numerator of the posterior is

$$p(ham|text) \propto p(ham) \prod_{w \in text} p(w|ham)^{N(w, text)}. \quad (8.7)$$

To avoid overflow, in practice we use log-probability:

$$\log p(ham|text) \propto \log p(ham) + \sum_{w \in text} N(w, text) \log p(w|ham). \quad (8.8)$$

To avoid zero probability for rare words (in that case the probability of the entire e-mail would be zero), instead of the fraction

$$p(w|ham) = \frac{\#\text{of times word } w \text{ used in } ham}{\text{total } \#\text{ of words in ham}} \quad (8.9)$$

8.4 Bayesian Classification

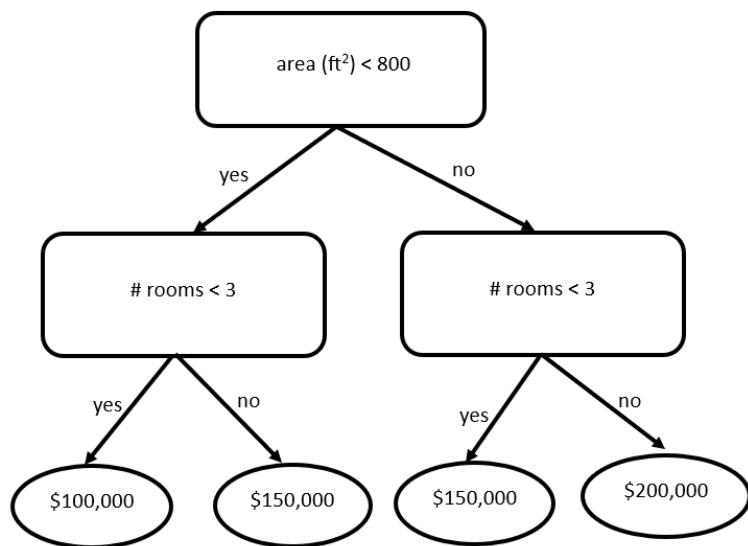
the following formula is used:

$$p(w|ham) = \frac{\text{#of times word } w \text{ used in ham} + \delta}{\text{total # of words in ham} + \delta|V|}. \quad (8.10)$$

9 Decision Trees

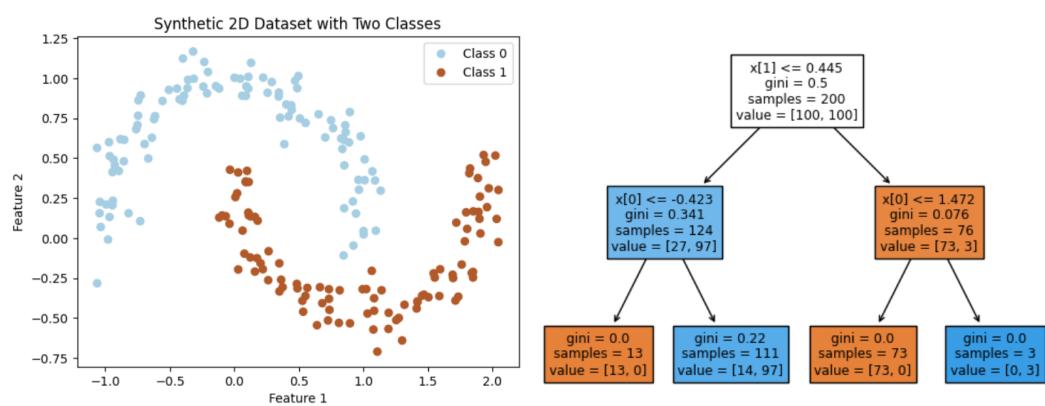
Creating new features

$$[700 < \text{area (ft}^2\text{)} < 800][1 < \#\text{of rooms} < 3]$$

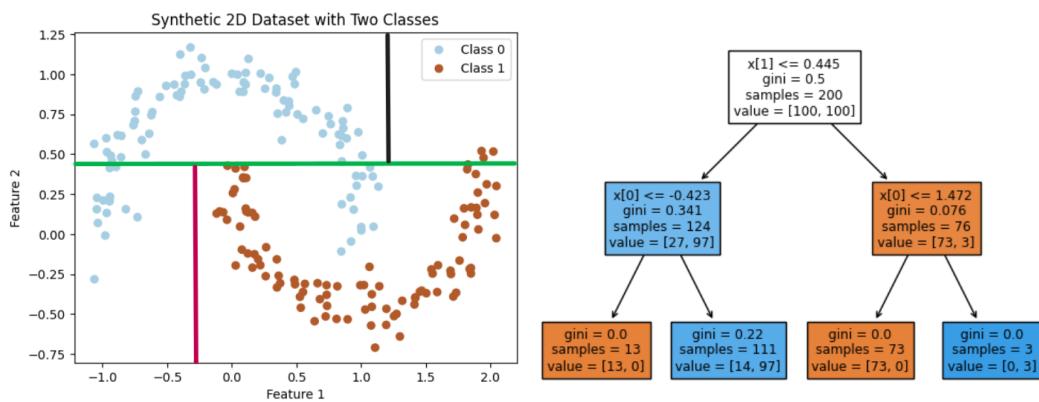
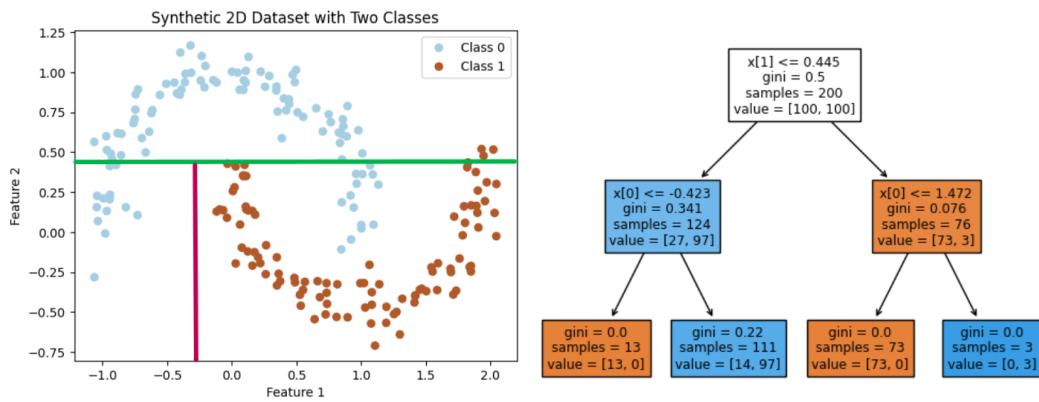
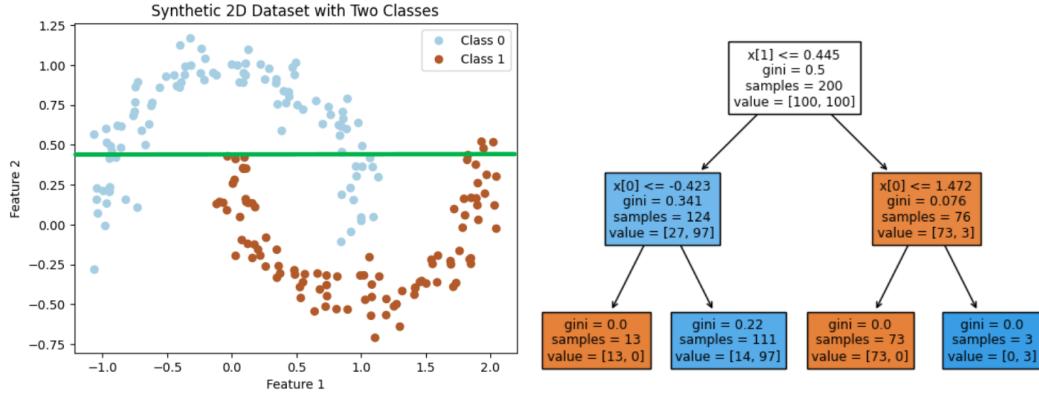




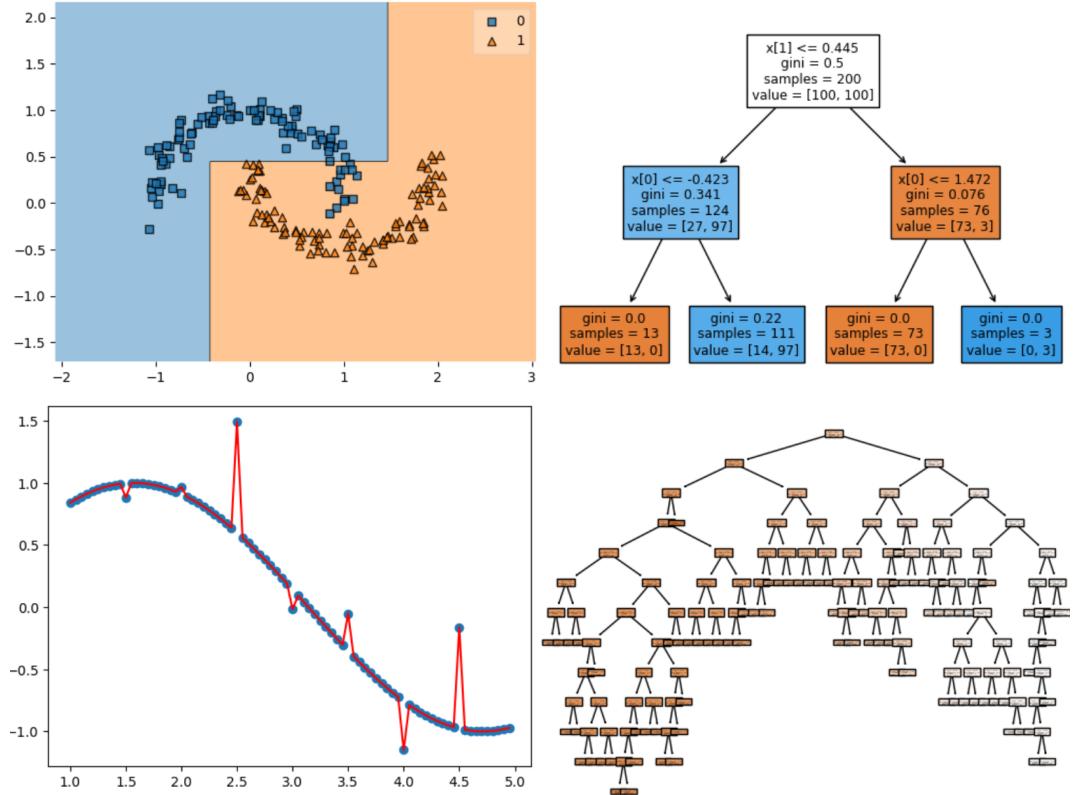
9.1 Geometry of Decision Tree Classifier



9.1 Geometry of Decision Tree Classifier



9 Decision Trees



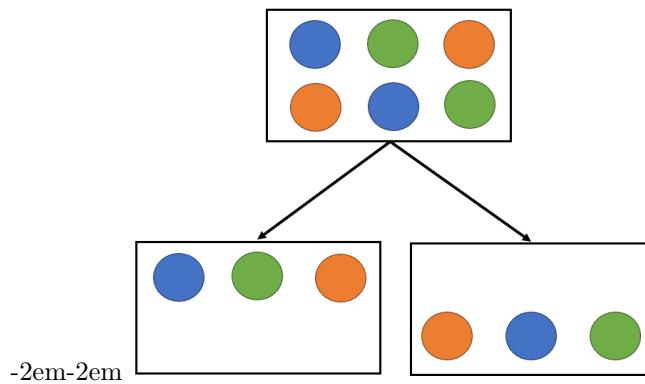
9.2 Leaf's Values

Classification:

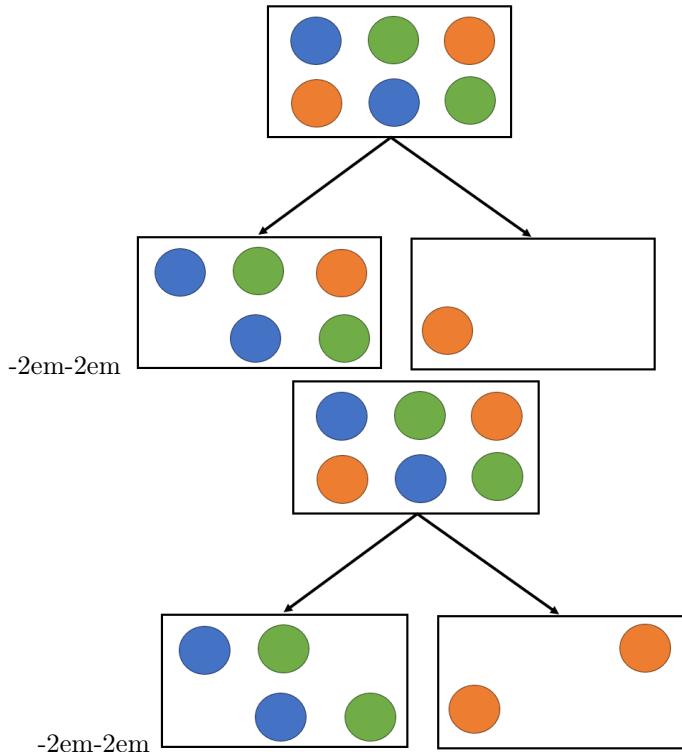
$$\operatorname{argmax}_c \sum_{x^{(i)} \in V} [y^{(i)} = c] = \operatorname{argmax}_c \frac{1}{|V|} \sum_{x^{(i)} \in V} [y^{(i)} = c].$$

Regression

$$\frac{1}{|V|} \sum_{x^{(i)} \in V} y^{(i)}$$



9.2.1 How to split one node? Classification



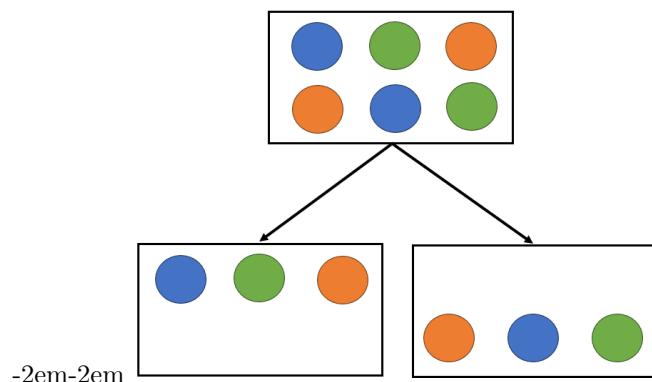
Information is defined by

$$I(p) = -\log p.$$

Entropy (work 'ergon' transformation 'tropy')

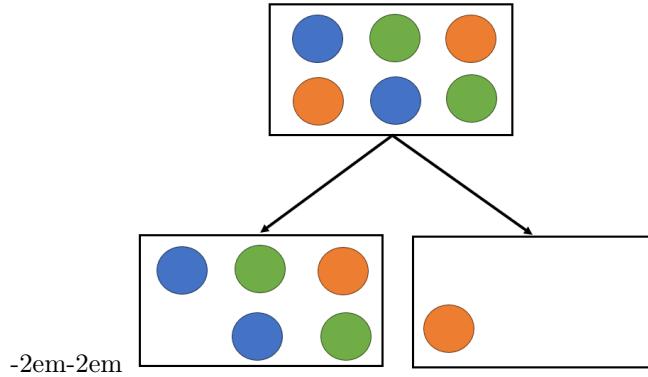
$$H(p) = E[I(p)] = - \sum_i p_i \log p_i.$$

9.2.2 Impurity



$$H(V) = -\frac{1}{3} \log \frac{1}{3} - \frac{1}{3} \log \frac{1}{3} - \frac{1}{3} \log \frac{1}{3} \approx 1.0986$$

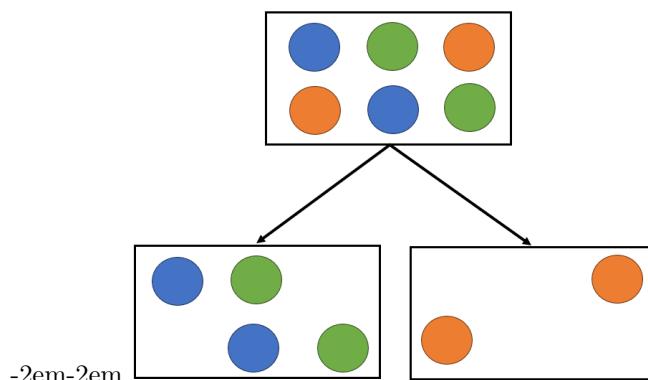
$$H(V_L) = H(V_R) = -\frac{1}{3} \log \frac{1}{3} - \frac{1}{3} \log \frac{1}{3} - \frac{1}{3} \log \frac{1}{3} \approx 1.0986$$



$$H(V) = -\frac{1}{3} \log \frac{1}{3} - \frac{1}{3} \log \frac{1}{3} - \frac{1}{3} \log \frac{1}{3} \approx 1.0986$$

$$H(V_L) = -\frac{2}{5} \log \frac{2}{5} - \frac{2}{5} \log \frac{2}{5} - \frac{1}{5} \log \frac{1}{5} \approx 1.055$$

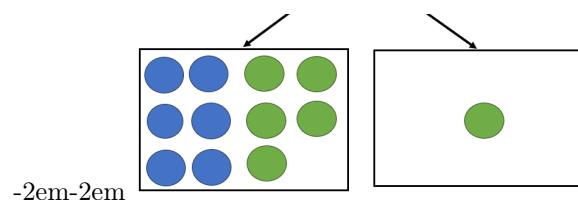
$$H(V_R) = 1 \log 1 = 0$$



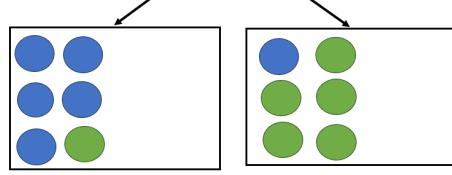
$$H(V) = -\frac{1}{3} \log \frac{1}{3} - \frac{1}{3} \log \frac{1}{3} - \frac{1}{3} \log \frac{1}{3} \approx 1.0986$$

$$H(V_L) = -\frac{1}{2} \log \frac{1}{2} - \frac{1}{2} \log \frac{1}{2} \approx 0.693$$

$$H(V_R) = 1 \log 1 = 0$$



$$H(V_L) = -\frac{6}{11} \log \frac{6}{11} - \frac{5}{11} \log \frac{5}{11} \approx 0.689, \quad H(V_R) = 1 \log 1 = 0.$$



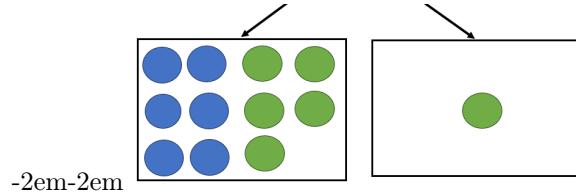
$$H(V_L) = -\frac{5}{6} \log \frac{5}{6} - \frac{1}{6} \log \frac{1}{6} \approx 0.45, \quad H(V_R) = H(V_L)$$

9.2.3 Impurity

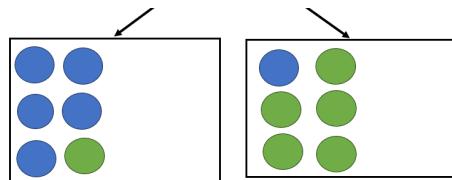
$$Q = H(V) - \left(\frac{|V_L|}{|V|} H(V_L) + \frac{|V_R|}{|V|} H(V_R) \right) \rightarrow \max$$

or

$$Q = \frac{|V_L|}{|V|} H(V_L) + \frac{|V_R|}{|V|} H(V_R) \rightarrow \min$$

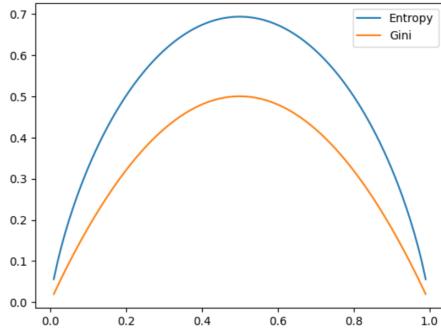


$$H(V_L) \approx 0.689, \quad H(V_R) = 1 \log 1 = 0, \quad Q = \frac{11}{12} H(V_L) + \frac{1}{12} H(V_R) \approx 0.632.$$



$$H(V_L) = H(V_R) \approx 0.45, \quad Q = \frac{1}{2} H(V_L) + \frac{1}{2} H(V_R) \approx 0.45.$$

9.2.4 Gini VS Entropy



In practice we use

$$H(V) = \sum_{i=1} p_i(1 - p_i)$$

instead of

$$H(V) = - \sum_{i=1} p_i \log p_i$$

9.3 Impurity for Regression

L_2

$$H(V) = \frac{1}{V} \sum_{x^{(i)} \in V} \left(y^{(i)} - \bar{y} \right)^2$$

or L_1

$$H(V) = \frac{1}{V} \sum_{x^{(i)} \in V} \left| y^{(i)} - \text{median}(y) \right|.$$

9.3.1 Decision Tree Formula

If we have leaves V_1, V_2, \dots, V_J , and a_1, a_2, \dots, a_J are the predictions in these nodes, then

$$a(x) = \sum_{j=1}^J a_j [x \in V_j].$$

10 Models Ensembles

10.1 Voting

10.2 Bootstrap Aggregation (Bagging)

10.3 Random Forest

10.3.1 Feature Importance

10.4 Boosting

10.4.1 Gradient Boosting Machine (GBM)

10.5 Stacking and Blending

11 Clustering

11.1 k -Means and k -Medoids

11.2 Hierarchical Clustering

11.3 Mixture of Gaussians

12 Dimensionality Reduction

12.1 PCA

12.1.1 Problem Formulation

In this notes, we derive PCA from the dimensionality reduction perspective.

We assume that we have n data points with p features:

$$x^{(i)} = \begin{bmatrix} x_1^{(i)} \\ \vdots \\ x_p^{(i)} \end{bmatrix}, \quad i = 1, \dots, n. \quad (12.1)$$

We want to find a low-dimensional representation of the data

$$z^{(i)} = \begin{bmatrix} z_1^{(i)} \\ \vdots \\ z_\ell^{(i)} \end{bmatrix}, \quad i = 1, \dots, n, \quad (12.2)$$

with a linear reconstruction

$$\hat{x}^{(i)} = Wz^{(i)}, \quad W \in M(p \times \ell). \quad (12.3)$$

We denote by $M(n_1 \times n_2)$ the space of real-valued matrices with n_1 rows and n_2 columns.

We want to minimize the reconstruction loss

$$L(W, z^{(1)}, \dots, z^{(n)}) = \sum_{i=1}^n \|x^{(i)} - \hat{x}^{(i)}\|^2 = \sum_{i=1}^n \|x^{(i)} - Wz^{(i)}\|^2 \quad (12.4)$$

under the additional assumption that the columns W_k of the matrix W are orthonormal:

$$\|W_i\|^2 = 1, \quad (W_i, W_j) = 0, \quad i, j = 1, \dots, \ell, \quad i \neq j. \quad (12.5)$$

Denote by X a matrix with vectors $x^{(i)}$ in rows:

$$X = \begin{bmatrix} -x^{(1)T} \\ -x^{(2)T} \\ \dots \\ -x^{(n)T} \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_p^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_p^{(2)} \\ \dots & \dots & \dots & \dots \\ x_1^{(n)} & x_2^{(n)} & \dots & x_p^{(n)} \end{bmatrix}. \quad (12.6)$$

Similarly, denote by Z a matrix with vectors $z^{(i)}$ in its rows. Then the loss function (12.4) can be written as follows¹:

$$L(W, Z) = \|X^T - WZ^T\|_F^2. \quad (12.7)$$

¹Recall that the Frobenius norm of a matrix A is the square root from the sum of squared elements of this matrix: $\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$

12 Dimensionality Reduction

Condition (12.5) is equivalent to the following equations:

$$W_i^T W_j = \delta_{ij}, \quad i, j = 1, \dots, \ell, \quad (12.8)$$

or in matrix notation

$$W^T W = I. \quad (12.9)$$

We arrive to the minimization problem:

$$\begin{cases} L(W, Z) = \|X^T - WZ^T\|_F^2 \rightarrow \min_{W, Z}, \\ W^T W = I. \end{cases} \quad (12.10)$$

Because it contains only equality constraints, we can use the Lagrange multipliers method. Compose the Lagrange function:

$$\mathcal{L}(W, Z, \Lambda) = \|X^T - WZ^T\|_F^2 - \sum_{i=1}^{\ell} \lambda_{ii} (W_i^T W_i - 1) - \sum_{\substack{i,j=1 \\ i < j}}^{\ell} \lambda_{ij} W_i^T W_j. \quad (12.11)$$

Fist, we take the derivatives² with respect to $z^{(i)}$:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial z^{(i)}} &= \frac{\partial}{\partial z^{(i)}} \sum_{k=1}^{\ell} \|x^{(k)} - Wz^{(k)}\|^2 \\ &= \frac{\partial}{\partial z^{(i)}} \sum_{k=1}^{\ell} (x^{(k)} - Wz^{(k)})^T (x^{(k)} - Wz^{(k)}) \\ &= \frac{\partial}{\partial z^{(i)}} \sum_{k=1}^{\ell} (x^{(k)T} - z^{(k)T} W^T) (x^{(k)} - Wz^{(k)}) \\ &= \frac{\partial}{\partial z^{(i)}} \sum_{k=1}^{\ell} (x^{(k)T} x^{(k)} - 2x^{(k)T} Wz^{(k)} + z^{(k)T} W^T Wz^{(k)}) \\ &= -2W^T x^{(i)} + 2z^{(i)} = 0. \end{aligned} \quad (12.12)$$

We find that

$$z^{(i)} = W^T x^{(i)}. \quad (12.13)$$

Substituting this representation in loss function (12.7), we obtain a new optimization problem:

$$\begin{cases} L(W) = L(W, Z(W)) = \|X^T - WW^T X^T\|_F^2 \rightarrow \min_W, \\ W^T W = I. \end{cases} \quad (12.14)$$

Rewrite the target function:³

$$\begin{aligned} \|X^T - WW^T X^T\|_F^2 &= Tr((X^T - WW^T X^T)^T (X^T - WW^T X^T)) \\ &= Tr((X - XWW^T)(X^T - WW^T X^T)) \\ &= Tr(XX^T - 2XWW^T X^T + XWW^T WW^T X^T) \\ &= Tr(XX^T - 2XWW^T X^T + XWW^T X^T) \\ &= Tr(XX^T - XWW^T X^T). \end{aligned} \quad (12.15)$$

²We use the following equalities $\frac{\partial a^T x}{\partial x} = a$ and $\frac{\partial x^T A x}{\partial x} = 2Ax$. See, e.g., https://en.wikipedia.org/wiki/Matrix_calculus

³We use the following identity $\|A\|_F^2 = Tr(A^T A)$. The trace of a matrix is defined as the sum of its diagonal elements.

Because the first term does not depend on W , problem (12.14) can be reformulated as follows:

$$\begin{cases} L(W) = \text{Tr}(-XWW^TX^T) \rightarrow \min_W, \\ W^TW = I. \end{cases} \quad (12.16)$$

Or equivalently⁴ as

$$\begin{cases} L(W) = \text{Tr}\left(\frac{1}{n}X^TXWW^T\right) \rightarrow \max_W, \\ W^TW = I. \end{cases} \quad (12.17)$$

The matrix $\frac{1}{n}X^TX$ is known as a covariance matrix and will be denoted by

$$\Sigma = \frac{1}{n}X^TX. \quad (12.18)$$

Compose the Lagrange function for this problem:

$$\mathcal{L}(W, \Lambda_{\ell \times \ell}) = \text{Tr}(\Sigma WW^T) - \sum_{i=1}^{\ell} \lambda_{ii}(W_i^TW_i - 1) - \sum_{\substack{i,j=1 \\ i < j}}^{\ell} \lambda_{ij}W_i^TW_j. \quad (12.19)$$

Here, we used a lower-triangular matrix $\Lambda_{\ell \times \ell} \in M(\ell \times \ell)$ to store the Laplace multipliers:

$$\Lambda_{\ell \times \ell} = \begin{bmatrix} \lambda_{11} & 0 & \cdots & 0 \\ \lambda_{21} & \lambda_{22} & \cdots & 0 \\ \dots & \dots & \dots & \dots \\ \lambda_{\ell 1} & \lambda_{\ell 2} & \cdots & \lambda_{\ell \ell} \end{bmatrix}. \quad (12.20)$$

Taking the derivatives with respect to columns W_k and λ_{ij} , we get the following system of equations:⁵

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial W_k} = 2\Sigma W_k - 2\lambda_{kk}W_k - \sum_{i=1}^{k-1} \lambda_{ik}W_i - \sum_{j=k+1}^{\ell} \lambda_{kj}W_j = 0, & k = 1, \dots, \ell, \\ \frac{\partial \mathcal{L}}{\partial \lambda_{kk}} = W_k^TW_k - 1 = 0, & k = 1, \dots, \ell, \\ \frac{\partial \mathcal{L}}{\partial \lambda_{ij}} = W_i^TW_j = 0, & i, j = 1, \dots, \ell, i < j. \end{cases} \quad (12.21)$$

If we multiply the first equation by W_k^T from the left, we will get

$$2W_k^T\Sigma W_k - 2\lambda_{kk}W_k^TW_k - \sum_{i=1}^{k-1} \lambda_{ik}W_k^TW_i - \sum_{j=k+1}^{\ell} \lambda_{kj}W_k^TW_j = 0, \quad k = 1, \dots, \ell. \quad (12.22)$$

Using the orthogonality conditions, we get

$$W_k^T\Sigma W_k - \lambda_{kk} = 0, \quad k = 1, \dots, \ell. \quad (12.23)$$

⁴We use the following property of the trace: $\text{Tr}(AB) = \text{Tr}(BA)$.

⁵To calculate the derivative $\frac{\partial \mathcal{L}}{\partial W_k}$ we use formula $\frac{\partial \text{Tr}(\Sigma WW^T)}{\partial W} = 2\Sigma W$ and then take k th column.

12 Dimensionality Reduction

And multiplying the last equation by W_k from the left, it takes form

$$\Sigma W_k = \lambda_{kk} W_k, \quad k = 1, \dots, \ell, \quad (12.24)$$

which is the eigenvalue problem for the matrix Σ .

Because the matrix $\Sigma \in M(p \times p)$ is symmetric,⁶ it has p orthogonal eigenvectors⁷ and corresponding eigenvalues are real and nonnegative.⁸ We can choose $\ell \leq p$ columns of the matrix W as normalized eigenvectors of the matrix Σ and λ_{kk} as corresponding eigenvalues. They satisfy equation (12.24) and conditions (12.9).

Due to the linear independence of the eigenvectors W_k from the first equation of (12.21) it follows that off-diagonal elements $\lambda_{ij} = 0$ with $i < j$. Therefore, matrix Λ has form

$$\Lambda_{\ell \times \ell} = \begin{bmatrix} \lambda_{11} & 0 & \cdots & 0 \\ 0 & \lambda_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_{\ell\ell} \end{bmatrix}. \quad (12.25)$$

If we substitute this solution to the target function, we will get

$$L(W) = \text{Tr}(\Sigma W W^T) = \text{Tr}(W^T \Sigma W) = \text{Tr}(\Lambda_{\ell \times \ell}) = \lambda_{11} + \dots + \lambda_{\ell\ell}. \quad (12.26)$$

To get the maximal value, we should use the first ℓ maximal eigenvalues of the matrix Σ .

12.1.2 Statistical Interpretation

Before applying PCA algorithm, i.e., calculating ℓ eigenvectors of the matrix Σ corresponding to the ℓ greatest eigenvalues, we it is recommended to subtract the mean:

$$x \rightarrow x - \bar{x}, \quad (12.27)$$

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x^{(i)} = \begin{bmatrix} \frac{1}{n} \sum_{i=1}^n x_1^{(i)} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n x_p^{(i)} \end{bmatrix}. \quad (12.28)$$

If we assume the zero mean of the points $x^{(i)}$, then the low-dimensional points

$$z^{(i)} = W^T x^{(i)} = \begin{bmatrix} -W_1^T - \\ \vdots \\ -W_\ell^T - \end{bmatrix} \begin{bmatrix} x_1^{(i)} \\ \vdots \\ x_p^{(i)} \end{bmatrix} = \begin{bmatrix} W_1^T x^{(i)} \\ \vdots \\ W_\ell^T x^{(i)} \end{bmatrix} \quad (12.29)$$

⁶For $A = X^T X$ $A^T = (X^T X)^T = X^T (X^T)^T = X^T X = A$.

⁷Theorem

⁸If h is an eigenvector of the matrix $X^T X$ with the eigenvalue λ , $X^T X h = \lambda h$. Therefore, $\lambda h^T h = h^T X^T X h = (Xh)^T (Xh) = \|Xh\|^2 \geq 0$. By definition of the eigenvector, $h \neq 0$, that means that $h^T h = \|h\|^2 > 0$; this leads to $\lambda \geq 0$.

will have zero mean as well. Then their variance can be calculated as follows:

$$\begin{aligned} S_{z_k}^2 &= \frac{1}{n} \sum_{i=1}^n (z_k^{(i)})^2 = \frac{1}{n} \sum_{i=1}^n (W_k^T x^{(i)})^2 = \frac{1}{n} \sum_{i=1}^n W_k^T x^{(i)} x^{(i)T} W_k \\ &= W_k^T \left(\frac{1}{n} \sum_{i=1}^n x^{(i)} x^{(i)T} \right) W_k = W_k^T \Sigma W_k = W_k^T \lambda_{kk} W_k = \lambda_{kk}. \end{aligned} \quad (12.30)$$

The total variance of the data points $z^{(i)}$ is equal to

$$\sum_{k=1}^n S_{z_k}^2 = \lambda_{11} + \dots + \lambda_{\ell\ell}. \quad (12.31)$$

The total variance of the data points $x^{(i)}$ is equal to⁹

$$\sum_{k=1}^n S_{x_k}^2 = \text{Tr}(\Sigma) = \text{Tr}(W_{p \times p} \Lambda W_{p \times p}^T) = \text{Tr}_{p \times p}(\Lambda) = \lambda_{11} + \dots + \lambda_{pp}. \quad (12.32)$$

This means that if we take $\ell = p$ components, then we reconstruct the data without any losses. It will correspond to the rotation of the original data. The fraction

$$\frac{\lambda_{11} + \dots + \lambda_{\ell\ell}}{\lambda_{11} + \dots + \lambda_{pp}} \quad (12.33)$$

represents the fraction of the explained variance.

12.1.3 Geometric Interpretation

We will show that reconstruction of the projection $z^{(i)}$ is the orthogonal projection of the point $x^{(i)}$ on the subspace spanned by vectors W_1, \dots, W_ℓ .

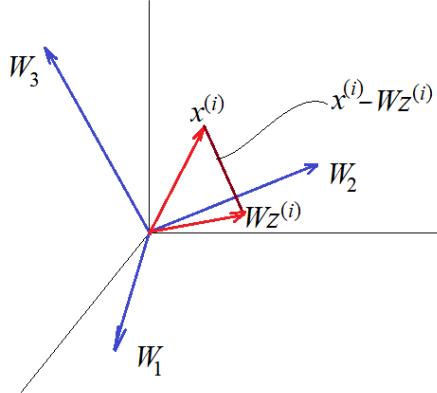


Figure 12.1: Orthogonal projection of a 3D vector $x^{(i)}$ on 2D plane of two principal components W_1 and W_2

It is sufficient to show that the residuals are orthogonal for each component:

$$x^{(i)} - W z^{(i)} \perp W_k, \quad k = 1, \dots, \ell. \quad (12.34)$$

⁹We assume that the eigenvalues are sorted in descending order.

12 Dimensionality Reduction

Equivalently, their dot product should be zero:

$$\begin{aligned}
W_k^T(x^{(i)} - Wz^{(i)}) &= W_k^T x^{(i)} - W_k^T \begin{bmatrix} | & & | \\ W_1 & \dots & W_\ell \\ | & & | \end{bmatrix} z^{(i)} \\
&= W_k^T x^{(i)} - [0, \dots, \underset{k}{1}, \dots, 0] z^{(i)} \\
&\stackrel{(12.29)}{=} W_k^T x^{(i)} - [0, \dots, \underset{k}{1}, \dots, 0] \begin{bmatrix} W_1^T x^{(i)} \\ \dots \\ W_\ell^T x^{(i)} \end{bmatrix} \\
&= W_k^T x^{(i)} - W_k^T x^{(i)} = 0.
\end{aligned} \tag{12.35}$$

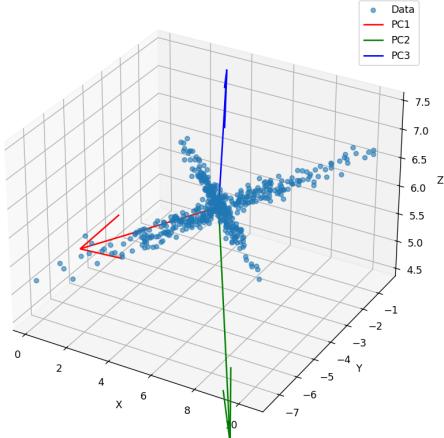
The explained variance ratio (12.33) provides another algorithm of PCA. Find the first component as a vector, such that projection of the data on it has the highest variance (see Fig. 12.2b):

$$\lambda_{11} = \max_{\|W_1\|=1} W_1^T \Sigma W_1. \tag{12.36}$$

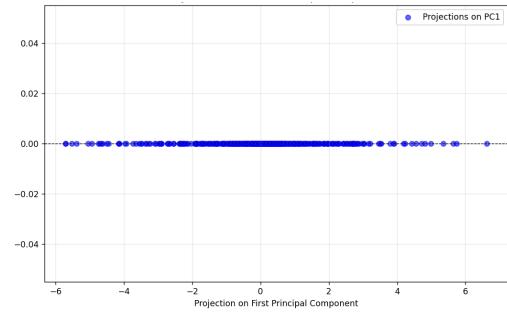
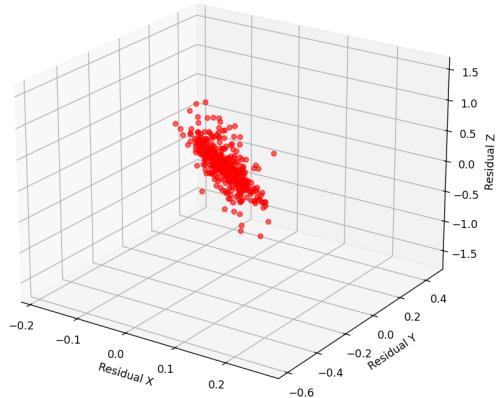
After obtaining the first component W_1 , we should find the second component orthogonal to W_1 , such that it captures the highest variance of the residuals (see Fig. 12.2c) $\tilde{x}^{(i)} = x^{(i)} - W_1 z^{(i)}$. This can be formulated as the following maximization problem:

$$\lambda_{22} = \max_{\|W_2\|=1} W_2^T \tilde{\Sigma} W_2, \tag{12.37}$$

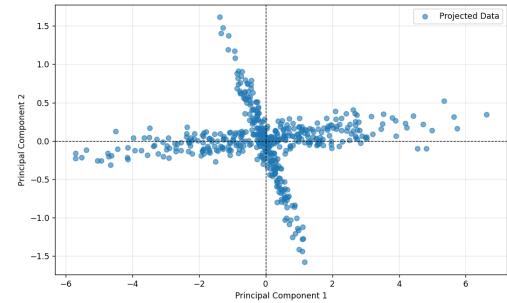
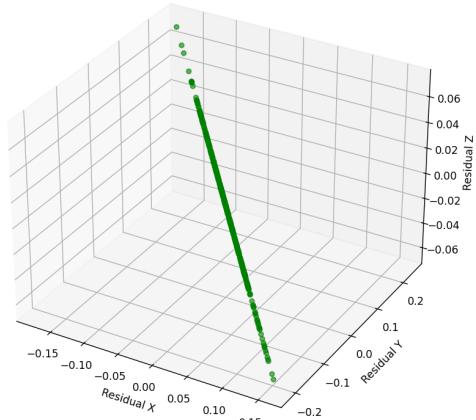
where $\tilde{\Sigma} = \frac{1}{n} \tilde{X}^T \tilde{X}$. And so on.



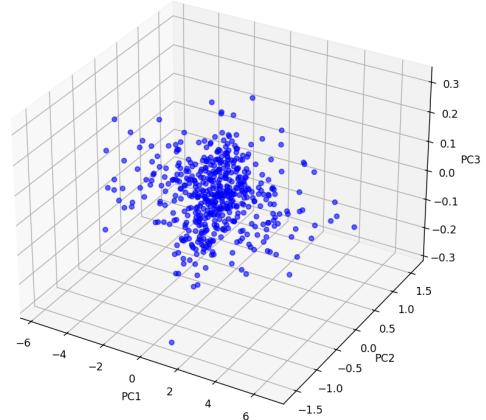
(a) Data with principal components.

(b) Projection on W_1 .

(c) Residuals lie on a plane.

(d) Projection on W_1 and W_2 .

(e) Residuals lie on a line.

(f) Projection on W_1 , W_2 , and W_3 .**Figure 12.2:** Illustration to sequential building of PCA.

12.1.4 Connection with Singular Value Decomposition (SVD)

Each matrix $A \in M(n_1 \times n_2)$ allows the singular value decomposition:

$$A_{n_1 \times n_2} = U_{n_1 \times n_1} \Sigma_{n_1 \times n_2} V^T_{n_2 \times n_2}, \quad (12.38)$$

with depending on $n_1 \geq n_2$ or $n_1 \leq n_2$

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sigma_{n_2} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \quad \text{or} \quad \Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sigma_{n_1} & 0 & \cdots & 0 \end{bmatrix}. \quad (12.39)$$

Where σ_i are eigenvalues of AA^T or A^TA . Columns of U are normalized eigenvectors of AA^T and columns of V are normalized eigenvectors of A^TA .

Consider the singular value decomposition of the matrix X :

$$X = USV^T. \quad (12.40)$$

Then

$$\frac{1}{n} X^T X = \frac{1}{n} V S U^T U S V^T = V \left(\frac{1}{\sqrt{n}} S \right)^2 V^T. \quad (12.41)$$

Comparison with

$$\frac{1}{n} X^T X = W \Lambda_{\ell \times \ell} W^T \quad (12.42)$$

demonstrates that variances can be calculated via squared singular values of the matrix X

$$\lambda_{ii} = \frac{\sigma_{ii}^2}{n}, \quad i = 1, \dots, \ell, \quad (12.43)$$

and the principle components W are its right singular vectors V .

12.2 t-Distributed Stochastic Neighbor Embedding (t-SNE)

In some cases PCA doesn't reflect the data structure. For example, in Fig. 12.3, the projections of the data on the first component overlap. To address this problem, we can use linear discriminant analysis (LDA), but for data visualization the t-distributed stochastic neighbor embedding algorithm (t-SNE) is typically used.

<https://jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf>

12.2.1 One Degree of Freedom in t-Distribution

Introduce the distribution on the discrete space of pairwise distances. The easiest way to do it is for the point $x^{(j)}$ to find k nearest neighbors (excluding itself)

$$x^{(j)}, x^{(j_1)}, \dots, x^{(j_k)}, \quad j_0 = j,$$

and define numbers

$$p_{j|j} = 0, \quad p_{j_1|j} = \dots = p_{j_k|j} = \frac{1}{k}, \quad p_{j_{k+1}|j} = \dots = p_{j_{n-1}|j} = 0. \quad (12.44)$$

12.2 *t*-Distributed Stochastic Neighbor Embedding (*t*-SNE)

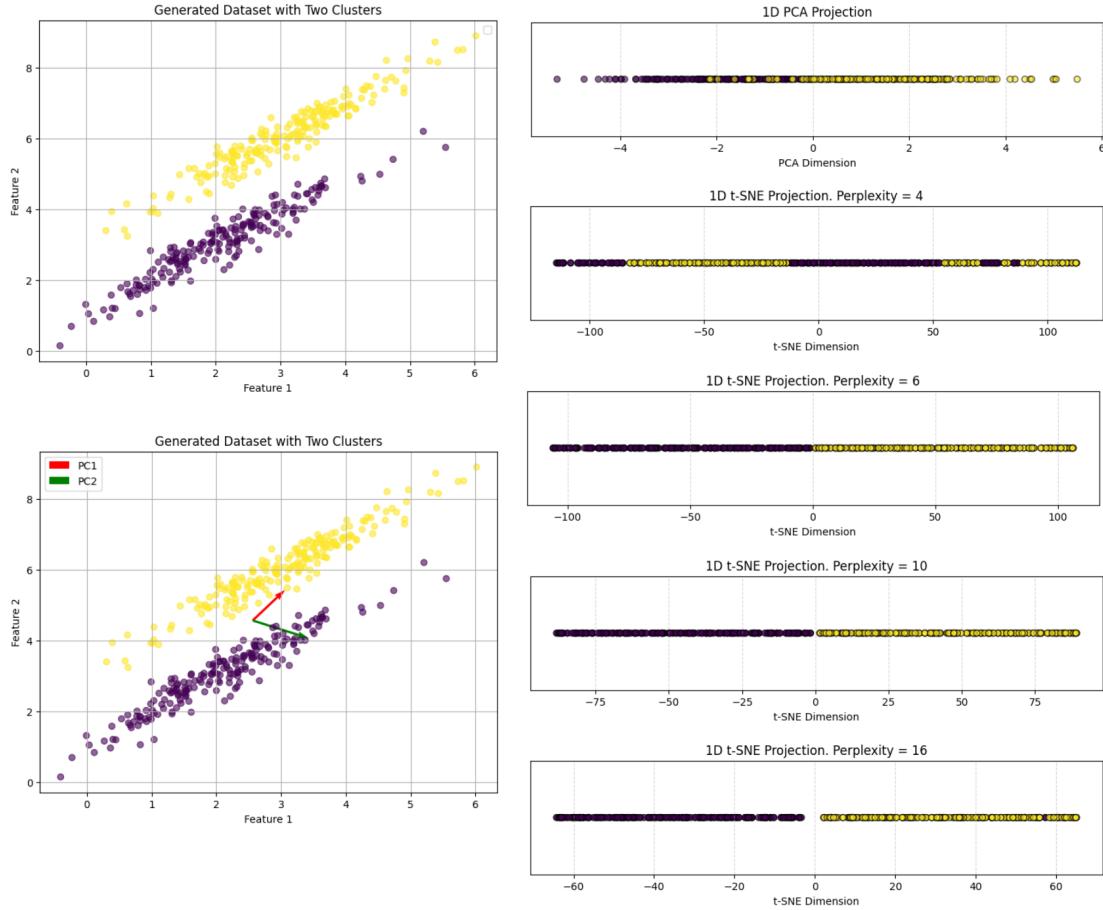


Figure 12.3: Comparison of PCA and t-SNE

Then for each pair $x^{(i)}$ and $x^{(j)}$ define probabilities

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2n}. \quad (12.45)$$

It is easy to check that

$$\sum_{i,j=1}^n p_{ij} = \frac{1}{2n} \left(\sum_{i=1}^n \sum_{j=1}^n p_{j|i} + \sum_{j=1}^n \sum_{i=1}^n p_{i|j} \right) = 1. \quad (12.46)$$

By *perplexity* we mean

$$\text{Perplexity} = 2^H, \quad (12.47)$$

where H is the Shannon's entropy.

For one point $x^{(j)}$ with probabilities () the perplexity is equal to the number of the neighbors:

$$\text{Perplexity} = 2^{-\sum_{i=1}^n p_{i|j} \log_2 p_{i|j}} = k. \quad (12.48)$$

12 Dimensionality Reduction

The common choice for the numbers $p_{i|j}$ is

$$p_{i|j} = \frac{e^{-\frac{|x^{(i)} - x^{(j)}|^2}{2\sigma_j^2}}}{\sum_{k=1}^n e^{-\frac{|x^{(k)} - x^{(j)}|^2}{2\sigma_k^2}}}, \quad (12.49)$$

where σ_j is chosen such that distribution $p_{i|j}$ has the desired perplexity.¹⁰ The standard choice of perplexity is 5 – 50.

The corresponding points $z^{(i)}$ in the lower-dimensional space are initialized randomly or using PCA.

The pairwise density is defined as follows:

$$q_{ij} = \frac{1}{Z} \frac{1}{1 + |z^{(i)} - z^{(j)}|^2}, \quad (12.50)$$

where Z is the normalizing constant:

$$Z = \sum_{\substack{i,j=1 \\ i \neq j}}^n \left(1 + |z^{(i)} - z^{(j)}|^2 \right)^{-1}. \quad (12.51)$$

The positions of the points $z^{(i)}$ are defined by minimization of the Kullback — Leibler divergence

$$KL(P||Q) = \sum_{r,s=1}^n p_{rs} \log \frac{p_{rs}}{q_{rs}}, \quad (12.52)$$

using gradient¹¹ descent:

$$z^{(i)} = z^{(i)} - \eta \nabla_{z^{(i)}} KL(P||Q), \quad i = 1, \dots, n. \quad (12.53)$$

To calculate the gradient with respect to $z^{(i)}$ denote by

$$d_{ij} = \|z^{(i)} - z^{(j)}\|^2 = z^{(i)T} z^{(i)} - 2z^{(i)T} z^{(j)} + z^{(j)T} z^{(j)} \quad (12.54)$$

and rewrite the minimizing function:

$$\begin{aligned} KL(P||Q) &= \sum_{r,s=1}^n p_{rs} \log p_{rs} - \sum_{r,s=1}^n p_{rs} \log q_{rs} \\ &= \sum_{r,s=1}^n p_{rs} \log p_{rs} - \sum_{r,s=1}^n p_{rs} \log(1 + d_{rs})^{-1} + \sum_{r,s=1}^n p_{rs} \log Z \\ &= \sum_{r,s=1}^n p_{rs} \log p_{rs} - \sum_{r,s=1}^n p_{rs} \log(1 + d_{rs})^{-1} + \log Z. \end{aligned} \quad (12.55)$$

¹⁰In practice, perplexity is chosen in the range 5—50.

¹¹By definition, the gradient is a vector of partial derivatives: $\nabla_{z^{(i)}} F(z^{(i)}) = \begin{bmatrix} \frac{\partial F}{\partial z_1^{(i)}} \\ \vdots \\ \frac{\partial F}{\partial z_\ell^{(i)}} \end{bmatrix}$. We use for it also notation of a matrix derivative $\frac{\partial F}{\partial z^{(i)}}$.

12.2 t-Distributed Stochastic Neighbor Embedding (t-SNE)

Then

$$\frac{\partial KL(P||Q)}{\partial z^{(k)}} = \sum_{i,j=1}^n \frac{\partial KL(P||Q)}{\partial d_{ij}} \frac{\partial d_{ij}}{\partial z^{(k)}}. \quad (12.56)$$

From (12.54) it follows that

$$\frac{\partial d_{ij}}{\partial z^{(k)}} = \begin{cases} 0, & i = j \text{ or } i \neq k, j \neq k, \\ 2z^{(k)} - 2z^{(j)}, & i = k, j \neq k, \\ -2z^{(i)} + 2z^{(k)}, & i \neq k, j = k. \end{cases} \quad (12.57)$$

From (12.56) it follows that

$$\begin{aligned} \frac{\partial KL(P||Q)}{\partial d_{ij}} &= \frac{p_{ij}}{1+d_{ij}} + \frac{1}{Z} \frac{\partial(1+d_{ij})^{-1}}{\partial d_{ij}} = p_{ij}(1+d_{ij})^{-1} - \frac{1}{Z}(1+d_{ij})^{-2} \\ &= \frac{p_{ij}(1+d_{ij})^{-1}}{Z} Z - q_{ij}(1+d_{ij})^{-1} = p_{ij}q_{ij}Z - q_{ij}^2Z \\ &= (p_{ij} - q_{ij})q_{ij}Z. \end{aligned} \quad (12.58)$$

$$\begin{aligned} \frac{\partial KL(P||Q)}{\partial z^{(k)}} &= \sum_{i \neq k} \frac{\partial KL(P||Q)}{\partial d_{ik}} (-2z^{(i)} + 2z^{(k)}) + \sum_{j \neq k} \frac{\partial KL(P||Q)}{\partial d_{kj}} (2z^{(k)} - 2z^{(j)}) \\ &= \sum_{i \neq k} (p_{ik} - q_{ik})q_{ik}Z (-2z^{(i)} + 2z^{(k)}) + \sum_{j \neq k} (p_{kj} - q_{kj})q_{kj}Z (2z^{(k)} - 2z^{(j)}) \\ &= 4 \sum_{i \neq k} (p_{ik} - q_{ik})q_{ik}Z (z^{(k)} - z^{(i)}). \end{aligned} \quad (12.59)$$

12.2.2 Other Degrees of Freedom

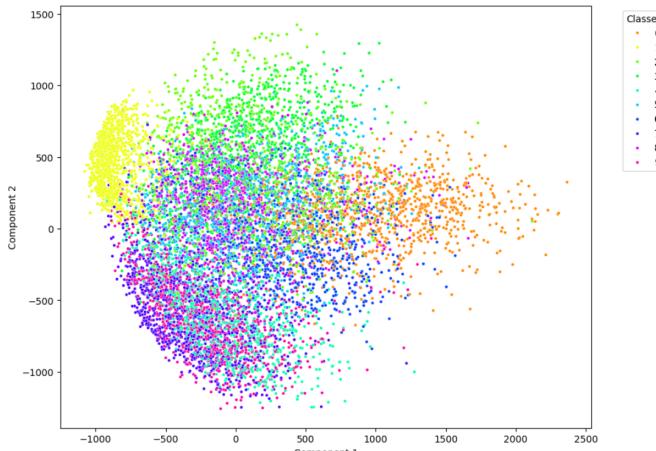
The pairwise density is defined as follows:

$$q_{ij} = \frac{1}{Z} \frac{1}{\left(1 + \frac{|z^{(i)} - z^{(j)}|^2}{\nu}\right)^{\frac{\nu+1}{2}}}, \quad (12.60)$$

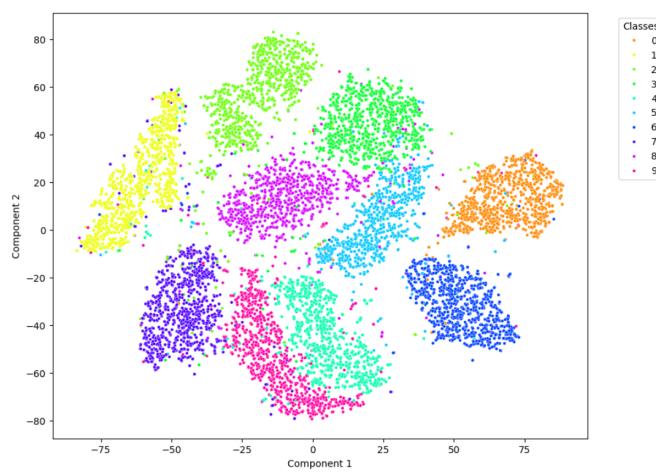
where Z is the normalizing constant:

$$Z = \sum_{\substack{i,j=1 \\ i \neq j}}^n \left(1 + \frac{|z^{(i)} - z^{(j)}|^2}{\nu}\right)^{-\frac{\nu+1}{2}}. \quad (12.61)$$

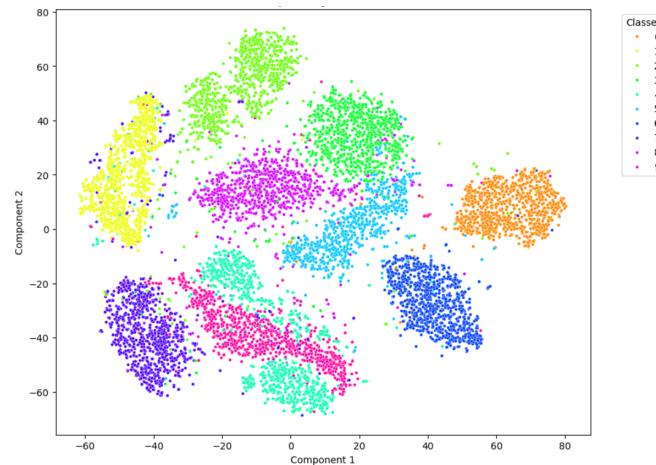
12 Dimensionality Reduction



(a) PCA.



(b) t-SNE with perplexity 30.



(c) t-SNE with perplexity 50.

Figure 12.4: Comparison of different dimensionality reduction algorithms on MNIST.

12.3 Orthogonal Procrustes Problem

This problem arises when we need to align two sets of points using only rotations and reflections. As in PCA we assume the data is centered.

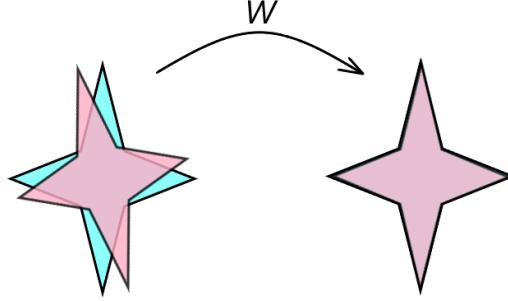


Figure 12.5: Illustration of the orthogonal alignment

Mathematically we want to find an orthogonal matrix W that maps a set of points X onto the points Y :

$$\begin{cases} L(W) = \sum_{i=1}^n \|Wx^{(i)} - y^{(i)}\|^2 = \|WX^T - Y^T\|_F^2 \rightarrow \min_W, \\ W^T W = I. \end{cases} \quad (12.62)$$

Rewrite the target function using trace:

$$\begin{aligned} L(W) &= \text{Tr}((WX^T - Y^T)^T(WX^T - Y^T)) \\ &= \text{Tr}((XW^T - Y)(WX^T - Y^T)) \\ &= \text{Tr}(XW^TWX^T - XW^TY^T - YWX^T + YY^T) \\ &= \text{Tr}(XX^T) - \text{Tr}(XW^TY^T) - \text{Tr}(YWX^T) + \text{Tr}(YY^T). \end{aligned} \quad (12.63)$$

Thus, the minimization problem (12.62) can be reformulated as a maximization problem:¹²

$$\begin{cases} L(W) = \text{Tr}(WX^T Y) \rightarrow \max_W, \\ W^T W = I. \end{cases} \quad (12.64)$$

If we consider SVD decomposition of the matrix $X^T Y$

$$X^T Y = U \Sigma V^T, \quad (12.65)$$

then¹³

$$\text{Tr}(WX^T Y) = \text{Tr}(WU\Sigma V^T) \leq \text{Tr}(V\Sigma V^T) = \text{Tr}(\Sigma). \quad (12.66)$$

Therefore, the matrix W^* delivering the maximum satisfies equation

$$W^* U = V \quad (12.67)$$

or

$$W^* = V U^T. \quad (12.68)$$

¹²We use the fact $\text{Tr}(A^T) = \sum_{i=1}^m a_{ii} = \text{Tr}(A)$ for a matrix $A \in M(m \times m)$.

¹³Theorem