COMP70015 Mathematics for Machine Learning

Timothy Chung

Autumn 2025

Contents

Chapter 0

Foreword

The notes are an amalgamation of the course's informal notes with some additions from the slides and any source material referenced.

0.1 Box Previews

Machine Learning — **Definition 0.1.1**

Machine learning is the study of computer algorithms that improve automatically through experience. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.

Machine Learning — ***Extra, Not Assessed* 0.1.1**

Machine learning is the study of computer algorithms that improve automatically through experience. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.

Machine Learning — **Example Question 0.1.1**

Machine learning is the study of computer algorithms that improve automatically through experience. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.

Q1c - 2018 — **Exam Question 0.1.1**

Machine learning is the study of computer algorithms that improve automatically through experience. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.

Machine Learning — **Author's Comment**

Machine learning is the study of computer algorithms that improve automatically through experience. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.

Machine Learning **Theorem 0.1.1**

Machine learning is the study of computer algorithms that improve automatically through experience. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.

Machine Learning **Reference 0.1.1**

Machine learning is the study of computer algorithms that improve automatically through experience. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.

Machine Learning **Intuition 0.1.1**

Machine learning is the study of computer algorithms that improve automatically through experience. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.

Chapter 1

Introduction and Formalising ML Problem Settings

1.1 Overview

Week 1: Discuss ML, brush up skills

Week 2: Critical optimisation concepts: automatic differentiation, convergence, complexity

Week 3: Probabilistic Perspective to ML

Week 4: Probabilistic Perspective to ML (contd.)

Week 5: Focus on Bayes Theorem + Practicals

Week 6: Consolidating learning (PCA study)

Week 7: Two Advanced topics: NN training with adversarial examples, + student selected topic

1.2 Brief Probability Theory Review

Rigour of A Random Variable **Author's Comment**

For the beginning of this module, a random variable X is considered to be a function from a sample space Ω to the unit interval $[0, 1]$.

However, in Lectures 6 and 7, we use the full definition of a random variable, which is a measurable function from a sample probability space (Ω) to a measurable space (E, ε), involcing the concept of σ-algebras and measure spaces.

When the sample space is a continuous space, the function p is a **probability density function** (PDF). When the sample space is a discrete space, the function p is a **probability mass function** (PMF).

$$x \sim X \tag{1.1}$$

x is a sample from a random variable X.

$$P(X = x) \tag{1.2}$$

This is the probability that the random variable X takes on the value x.

1.2.1 Probability Density Function (PDF)

The Probability Density Function (PDF) of a continuous random variable X is a function $f_X(x)$ that describes the relative likelihood for this random variable to take on a given value. The PDF has the following properties:

- $f_X(x) \geq 0$ for all x.

- $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$.

Mathematically, the PDF is defined such that the probability that X lies within a particular interval $[a, b]$ is given by:

$$P(a \leq X \leq b) = \int_a^b f_X(x)\,dx.$$

Example of a PDF: Normal Distribution

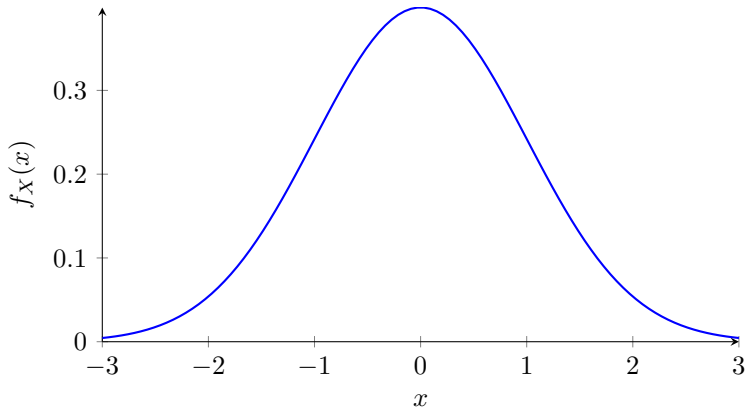


Figure 1.1: Probability Density Function of a standard normal distribution

1.2.2 Cumulative Distribution Function (CDF)

The Cumulative Distribution Function (CDF) of a continuous random variable X is a function $F_X(x)$ that describes the probability that X will take a value less than or equal to x. It is defined as:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\,dt.$$

The CDF has the following properties:

- $0 \leq F_X(x) \leq 1$ for all x.
- $F_X(x)$ is a non-decreasing function.
- $\lim_{x \to -\infty} F_X(x) = 0$.
- $\lim_{x \to \infty} F_X(x) = 1$.

Example of a CDF: Sigmoid Approximation

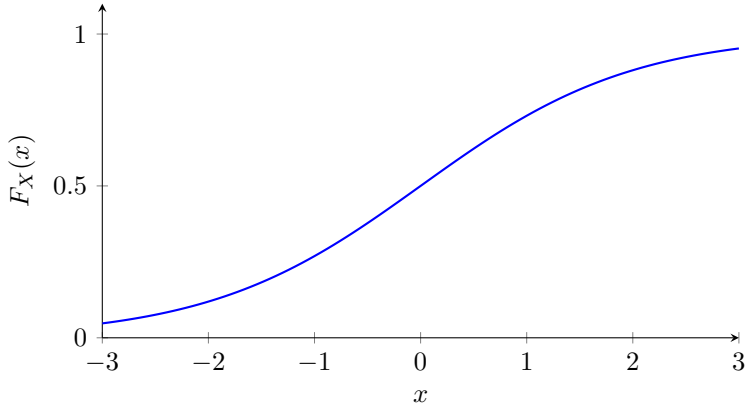


Figure 1.2: Cumulative Distribution Function

1.2.3 Dealing with Data

In Machine Learning (ML), we often deal with large datasets from which we aim to extract meaningful insights. It is useful to model these data points as random variables:

$$\{x_1, x_2, \ldots, x_N\}$$

We typically model these samples as random variables:

$$\{x_1 \sim X_1, x_2 \sim X_2, \ldots, x_N \sim X_N\}$$

1.2.4 Vectors of Random Variables

When dealing with joint random variables, we often represent them as vectors:

$$\mathbf{X} = (X_1, X_2, X_3) \quad \text{with joint distribution} \quad p_{X_1,X_2,X_3}(x_1, x_2, x_3)$$

For a vector of random variables $\mathbf{X}$ in $\mathbb{R}^3$, the joint probability density function is denoted as:

$$p_{\mathbf{X}}(\mathbf{x}) \quad \text{where} \quad \mathbf{X} \in \mathbb{R}^3$$

If these are input observations, they are often referred to as "feature vectors."

1.2.5 Distinct Random Variables

We model samples as distinct random variables:

$$\{x_1 \sim X_1, x_2 \sim X_2, \ldots, x_N \sim X_N\}$$

Question: Why should we model these as distinct random variables? Aren't they all the same thing?
Answer: Despite being identically distributed, distinct random variables allow for capturing dependencies and interactions between different samples.

1.3 Independent and Identically Distributed (i.i.d.) Random Variables

1.3.1 Independence of Random Variables

Independence refers to the idea that the values or outcomes of one observation in a dataset do not depend on or influence the values of any other observation.

For the set of random variables $X_1, X_2, \ldots X_n$, for the collection $\{x_1 \sim X_1, x_2 \sim X_2, \ldots, x_N \sim X_N\}$, we often assume independence, for any subset of observations $\{X_{i1}, X_{i2}, \ldots, X_{ik}\}$ where $i_1, i_2, \ldots, i_k \in \{1, 2, \ldots, N\}$, the joint distribution factorises:

$$P(X_{i_1} = x_{i_1}, \ldots, X_{i_k} = x_{i_k}) = P(X_{i_1} = x_{i_1}) \cdot \ldots \cdot P(X_{i_k} = x_{i_k}) \tag{1.3}$$

The joint distribution of the subset is the product of the marginal distributions of the individual random variables.

Independence: The occurrence of any event does not affect the occurrence of others. This is a crucial assumption in many ML models that we will see the usefulness of as early as the next lecture.

1.3.2 Identically Distributed Random Variables

The "identically distributed" part of i.i.d. refers to the idea that all random variables in the sample follow the same probability distribution. That is, they share the same probability density function (pdf) or probability mass function (pmf), depending on whether the data is continuous or discrete. If the set of random variables $\{X_1, X_2, \ldots, X_n\}$ are identically distributed, then:

$$f_{X_1}(x) = f_{X_2}(x) = \ldots = f_{X_n}(x) = f_X(x) \tag{1.4}$$

$$F_{X_1}(x) = F_{X_2}(x) = \ldots = F_{X_n}(x) = F_X(x) \tag{1.5}$$

This implies that the cumulative distribution function (CDF) is the same for all these random variables.

1.4 Statistical Modelling as Curve Fitting

Machine learning and statistics have significant overlap, and so most concepts studied in this module can be cast as statistical modelling. Consider our random variables:

$$\{x_1 \sim X_1, x_2 \sim X_2, \ldots, x_N \sim X_N\}$$

These random variables can be modelled to fit certain curves that represent the underlying data distribution.

1.4.1 Assumption about the Model

To proceed with statistical modelling, we make an assumption about the form of the model:

$$P_X(x) \approx \mathcal{N}(x; \theta) \quad \theta := (\mu, \sigma)$$

Here, $P_X(x)$ is approximated by a normal distribution $\mathcal{N}(x; \theta)$, where θ represents the parameters of the distribution, namely the mean μ and the standard deviation σ. This assumption simplifies the process of modelling the data.

1.4.2 Fitting the Model

The next step is to fit the model to the data. This involves finding the parameter values θ that maximize the probability of observing the given data. Mathematically, this is expressed as:

$$\arg\max_{\theta} P(x_1, \ldots, x_N \mid \theta)$$

In other words, we adjust the parameters μ and σ so that the assumed model best fits the observed data. This process is known as maximum likelihood estimation (MLE).

The Assumption of (i.i.d.) – Why? **Author's Comment**

Why do we model samples as distinct random variables?

1. **Variability:** Even though samples may be identically distributed, they are not the same. Each sample can take different values, and by modelling them as distinct random variables, we can account for this variability.
2. **Dependence Structure:** By treating samples as distinct, we can study and model the relationships and dependencies between them. This is important to understand the overall behavior of the system and in developing robust machine learning models.
3. **Central Limit Theorem:** Many results in probability and statistics, such as the Central Limit Theorem, rely on the assumption that we have a collection of distinct, identically distributed random variables.
4. **Learning and Generalisation:** Assuming independence simplifies the mathematical analysis and derivation, and to use techniques like maximum likelihood estimation and empirical risk minimisation. Independence ensures a model trained on a subset of data can generalise well to unseen data. If samples were not independent, the model would overfit to specific data points that do not hold in general.

1.5 Probability Density Function (PDF) and Cumulative Distribution Function (CDF)

1.5.1 Probability Density Function (PDF)

The Probability Density Function (PDF) $f_X(x)$ describes the relative likelihood for a continuous random variable X to take on a given value x:

$$\int_{-\infty}^{\infty} f_X(x)\, dx = 1$$

For any interval $[a, b]$:

$$P(a \leq X \leq b) = \int_a^b f_X(x)\,dx$$

Example of a PDF: Normal Distribution

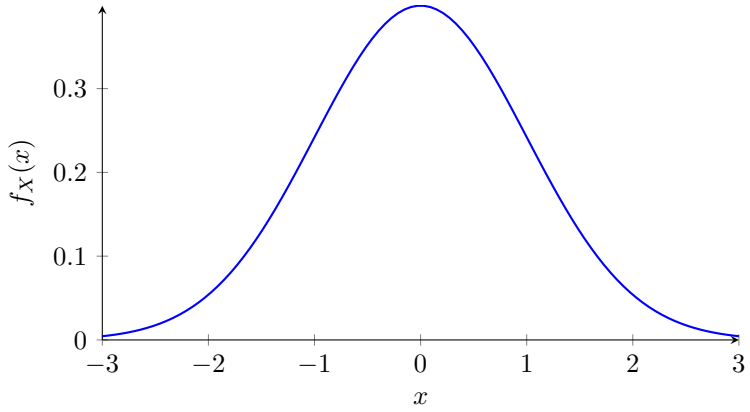


Figure 1.3: Probability Density Function of a standard normal distribution

1.5.2 Cumulative Distribution Function (CDF)

The Cumulative Distribution Function (CDF) $F_X(x)$ describes the probability that a continuous random variable X will take a value less than or equal to x:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\,dt$$

Properties of CDF:

- $0 \leq F_X(x) \leq 1$ for all x
- $F_X(x)$ is a non-decreasing function
- $\lim_{x \to -\infty} F_X(x) = 0$
- $\lim_{x \to \infty} F_X(x) = 1$

Example of a CDF: Sigmoid Approximation

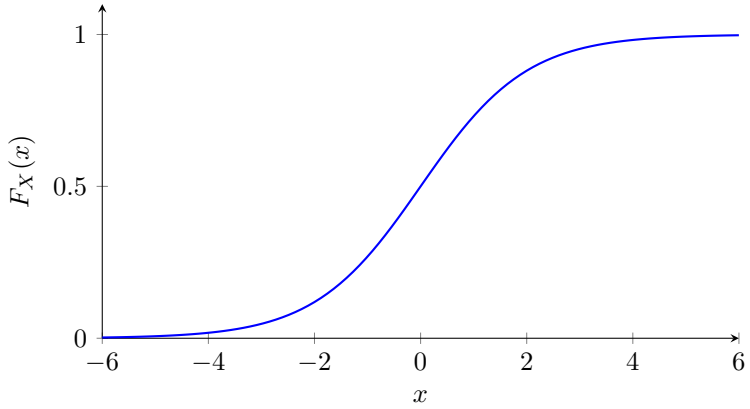


Figure 1.4: Cumulative Distribution Function (Sigmoid Approximation)

1.6 Discrete Random Variables

1.6.1 Bernoulli and Multinoulli Distributions

Bernoulli Distribution: Outcome with two values (e.g., heads or tails).

- Parameter: θ.
- Probabilities: $P(X = 0) = 1 - \theta$, $P(X = 1) = \theta$.

Multinoulli Distribution: Describes a scenario with multiple possible outcomes, extending the Bernoulli distribution to more than two outcomes.

- Parameter: $\boldsymbol{\theta} \in \mathbb{R}^s$ where each θ_i represents the probability of the i-th outcome, and $\sum_{i=0}^{s-1} \theta_i = 1$.
- Probabilities: $P(X = i) = \theta_i$ for $i = 0, 1, \ldots, s - 1$.

1.6.2 Binomial and Multinomial Distributions

Binomial Distribution: $X \sim \text{Bin}(n, \theta)$.

- Probability Mass Function (p.m.f):

$$\text{Bin}(k \mid n, \theta) := \binom{n}{k} \theta^k (1 - \theta)^{n-k}$$

- Binomial Coefficient:

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

- Mean: $n\theta$.
- Variance: $n\theta(1 - \theta)$.

Multinomial Distribution: Generalises the binomial distribution for more than two outcomes. It models the probabilities of counts among multiple categories in n independent trials.

- Parameters: n (number of trials) and $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$ where θ_i is the probability of the i-th category and $\sum_{i=1}^{K} \theta_i = 1$.
- Probability Mass Function (p.m.f):

$$\text{Mu}(\mathbf{x} \mid n, \boldsymbol{\theta}) := \binom{n}{x_1, x_2, \ldots, x_K} \prod_{j=1}^{K} \theta_j^{x_j}$$

  where $\mathbf{x} = (x_1, x_2, \ldots, x_K)$ represents the count of occurrences for each category.
- Multinomial Coefficient:

$$\binom{n}{x_1, x_2, \ldots, x_K} = \frac{n!}{x_1! x_2! \cdots x_K!}$$

- Mean for each category i: $\mathbb{E}[X_i] = n\theta_i$.
- Variance for each category i: $\text{Var}(X_i) = n\theta_i(1 - \theta_i)$.
- Covariance between categories i and j: $\text{Cov}(X_i, X_j) = -n\theta_i\theta_j$.

1.6.3 Empirical Distribution

Empirical Distribution **Intuition 1.6.1**

Definition of Empirical: Based on observation or experience rather than theory or pure logic.

Practically, we do not have access to an infinite amount of data, but we have instead a small fraction of it, a sample, to infer any insights from it. In the case of discrete random variables, we use probability mass functions, which is straightforward, but we are interested in probability density functions for continuous random variables, because to model the true distribution, we would need an infinite number of samples.

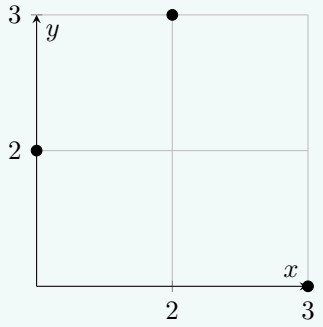
Thus, our goal is to approximate the true PDF from a given data set using finite samples. The transformation from discrete to continuous is done with the Dirac delta function.

A Dirac delta function is interestingly helpful because:

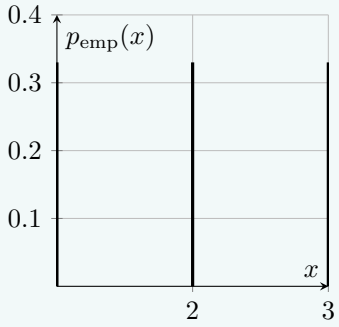
$$\int_{-\infty}^{\infty} \delta(x)\, dx = 1$$

We can then have multiple Dirac delta functions to represent the empirical distribution of a data set, but scaled down by a factor equivalent to the total number of data points to ensure the area under the curve is 1.

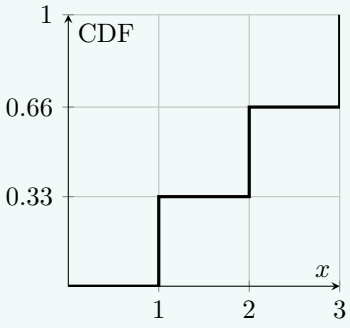
Plot 1: Data Points on a 2D Plane



Plot 2: Dirac Delta Functions



Plot 3: Cumulative Distribution Function (CDF)



Recommended Viewing **Author's Comment**

Source:
Empirical Statistics: When you compute statistics from a dataset, you're really computing statistics for its empirical distribution.

A dataset is a distribution.

Empirical Distribution: Suppose we have a set of data samples $D = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$. dervied from a random variable X. We can approximate the distribution of X using a set of delta functions on these samples:

- For a given data set D, the empirical distribution $p_{\text{emp}}(x)$ is defined as:

$$p_{\text{emp}}(x) := \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}(x)$$

  where $\delta_{x_i}(x)$ is the Dirac measure centred at x_i.
- **Dirac Measure:** $\delta_{x_i}(x)$ is a function that is 1 if $x = x_i$ and 0 otherwise. Formally, it is defined as:

$$\delta_{x_i}(x) = \begin{cases} 1, & \text{if } x = x_i \\ 0, & \text{if } x \neq x_i \end{cases}$$

- Explanation: The empirical distribution assigns equal probability $\frac{1}{N}$ to each observed data point x_i. It is a discrete distribution that places mass only on the observed data points.
- In general, one can associate weights with each element of the empirical distribution, i.e., $p_{\text{emp}}(x) = \frac{1}{N} \sum_{i=1}^{N} w_i \delta_{x_i}(x)$ as long as each $0 \leq w_i \leq 1$ and $\sum_{i=1}^{N} w_i = 1$.

1.7 Continuous Random Variables

1.7.1 Gaussian (Normal) Distribution

Probability Density Function (p.d.f):

- Formula:

$$N(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Mean: $\mu = \mathbb{E}[X]$.
- Variance: $\sigma^2 = \text{Var}[X]$.
- Standard Normal Distribution: $X \sim N(0, 1)$.
- Precision: $\lambda = \frac{1}{\sigma^2}$.

Cumulative Distribution Function (CDF):

- Formula:

$$\Phi(x; \mu, \sigma^2) = \int_{-\infty}^{x} N(z; \mu, \sigma^2)\, dz$$

- In terms of the error function (erf):

$$\Phi(x; \mu, \sigma^2) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{z}{\sqrt{2}}\right)\right]$$

  where $z = \frac{x-\mu}{\sigma}$ and

$$\text{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{0}^{x} \exp(-t^2)\, dt$$

1.8 Example Scene: Self-Driving Car

Notation and Gotchas **Author's Comment**

The lecture use x_i to denote the i-th feature, and $x^{(i)}$ to denote the i-th data point. The slides however seem to denote x_i as the i-th data point. The notes follow the former convention.

Also, it is generally assumed that in a classification problem, the classes are distinct and mutually exclusive, so an image is either a dog or a cat (multi-class classfication) but not a combination of both (multi-label classification). The course focuses on the former convention.

Take a case of the self-driving car. There are several facets:

1. **Object Recognition:** identifying objects in the scene

   - An image is a 2D array of pixel values. For a colour image, each pixel has 3 values (RGB).
   - **Input:**
   $$x \in \mathbb{R}^{H \times W \times C} \tag{1.6}$$
   where H is height, W is width, and C is the number of channels (e.g. 3 for RGB)
   - **Output:** We would like to detect if something is a background, another vehicle, the ground level, or a pedestrian. Let's say there are m outcomes, making this a classification problem. Naively, let $y \in \mathbb{R}^m$. However, these numbers are just arbitrary i.e. raw scores. It would make more sense to refine the output to a probability distribution over the m classes:
   $$y \in \Delta^m \quad \text{where} \sum_{i=1}^{m} y_i = 1 \tag{1.7}$$
   Where the Δ^m is the m-simplex, where the sum of all elements is 1. This provides a measure of "confidence" in the prediction. Example: To choose between four classes: [Vehicle, Pedestrian, Road, Background], the output could be $[0.1, 0.7, 0.1, 0.1]$. This means the model is 70% confident that the object is a pedestrian.
   - **Our objective is to learn a function:**
   $$f^\theta : x^{(i)} \mapsto y^{(i)} \quad \text{where } \theta \text{ are model parameters}, x^{(i)} \in \mathbb{R}^{H \times W \times C}, y^{(i)} \in \Delta^m \tag{1.8}$$

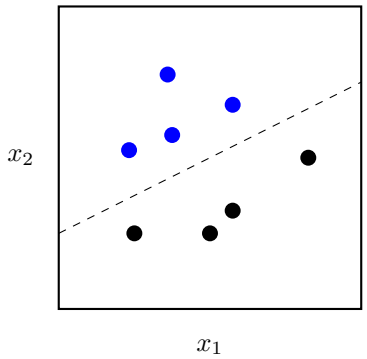
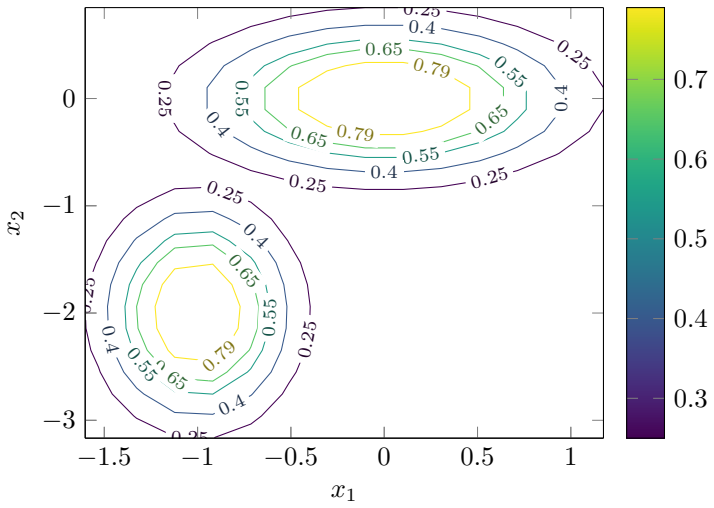


Figure 1.5: Simplified example for f^θ

   A simplified example for f^θ is shown in Figure 1.5. Ideally, we want a separator that separates the input space into m distinct regions according to observed data.
   - **Data Formalisation (1):** We can view our dataset as a set of N pairs where $x^{(i)}$ is an image and $y^{(i)}$ is the corresponding label. More neatly put,
   $$\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}, \quad x^{(i)} \in \mathbb{R}^{H \times W \times C}, y^{(i)} \in \Delta^m \tag{1.9}$$
   - **Data Formalisation (2):** We could also view the dataset as an empirical distribution over the data space:
   $$\mathcal{D} = \{x^{(i)}, y^{(i)}\}_{i=1}^{N} \sim p_{\text{data}}(x, y) \tag{1.10}$$



2. Route planning: deciding where to go
3. Speed Control: deciding how fast to go
4. Steering Angle Prediction: deciding how to steer

1.9 Mathematics of Linear Models

Linear models are fundamental in statistics and supervised machine learning. They can:

- Handle linear relationships.
- Be augmented with kernels or basis functions to model non-linear relationships.
- Provide analytical tractability for studying concepts like convergence, probabilistic modelling, and overfitting.

This section introduces the linear regression model.

1.9.1 Linear Regression Model

Linear regression involves performing regression with a linear model in a supervised setting. Given a dataset $\mathcal{D}$ consisting of input-output pairs $(\boldsymbol{x}^{(i)}, y^{(i)})$ for $i = 1, \ldots, N$:

- $\boldsymbol{x} \in \mathbb{R}^n$ represents the input features.
- $y \in \mathbb{R}$ represents the output or target variable.
- Assume a domain $\mathbb{R}^n$ and a one-dimensional co-domain.
- Model: $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\theta}$.
- Noise: $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

The full model is:

$$\hat{y}^{(i)} = \boldsymbol{x}^{(i)\top} \boldsymbol{\theta} + \epsilon$$

The objective is to find $\boldsymbol{\theta}$ such that $\hat{y}^{(i)} \approx y^{(i)}$. In other words, we want to find a set of parameters $\boldsymbol{\theta}$ that best explains the relationship between the input features $\boldsymbol{x}$ and the target variable y. This means minimising the difference between the predicted values $\hat{y}$ and the actual values y.

Conventions and Notations:

- Vectors $\boldsymbol{x} \in \mathbb{R}^n$ are column vectors, written as $n \times 1$ matrices.
- $\boldsymbol{x}^\top$ (x transpose) swaps rows and columns of $\boldsymbol{x}$, resulting in a $1 \times n$ row vector.

Lecture Reading **Reference 1.9.1**

Chapter 1-2 of Kevin Murphy's *Machine Learning: A Probabilistic Perspective*. Chapter 1.4 was not covered, but may be of interest