Machine Learning and Data Mining

Jimmy Tsz Ming Yue[1]

University of Sydney

jyue6728@uni.sydney.edu.au

[1] 440159151

Contents

Chapter 1

Introduction to Machine Learning

1.1 What is Machine Learning?

> **Definition. Machine Learning**: Construction of Statistical models that is an underlying distribution from which the data is drawn, or using which we classify data into different categories.

1.1.1 Problems in Machine Learning

1. Prediction, involving classification and regression
2. Clustering, segmentation and summarisation which seeks to find pattterns in data.
3. Outlier and anamoly detection, which seeks to find unusual patterns.

a

1.1.2 Data Representation

In the field of Machine Learning there is a variety of data that needs to be interpreted. For a computer to make sense of such data, it must be in a form recognizable to it. Such a data form is a matrix which allows a variety of mappings to occur mathematically, which is desirable for Machine Learning Algorithms.

Similarity

A special form of a matrix is called the vector, which is a mathematical object belonging in a vector space.

> **Definition. Similarity**: Given two vectors $\vec{x}$ and $\vec{y}$, the cosine similarity is given by:
>
> $$\text{sim}(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \, \|\vec{y}\|} \tag{1.1}$$

The cosine similarity has various properties which we list:

1. $\text{sim}(\vec{x}, \vec{y}) = 0$ when $\vec{x} \cdot \vec{y} = 0$ which is when $\vec{x}$ is orthogonal to $\vec{y}$

2. $\text{sim}(\vec{x}, \vec{y}) = 1$ when $\vec{x}$ parallel to $\vec{y}$
3. $\text{sim}(\vec{x}, \vec{y}) < 0$ when vectors $\vec{x}$ and $\vec{y}$ are diametrically opposed.
4. $\text{sim}(\vec{x}, \vec{y})$ can never be larger than 1 due to the Cauchy-Schwarz Inequality:

Cauchy Schwarz Inequality For all $\vec{u}$ and $\vec{v}$ in an inner product space then; (in the case of the Euclidean inner product):

$$|\vec{x} \cdot \vec{y}| \leq \|\vec{x}\| \, \|\vec{y}\| \tag{1.2}$$

Chapter 2

Linear Algebra and Matrix Decompositions

2.1 Vector Spaces

Definition. A vector space V is defined to be a set that is closed under scalar multiplication and vector addition. This closure gives rises to the following axioms, let $\mathcal{F}$ define the field of scalars.

1. Associativity of vector addition:

$$(\vec{u} + \vec{v}) + \vec{w} = \vec{u} + (\vec{v} + \vec{w})$$

2. Commutavity of vector addition:

$$u + v = v + u, \quad \forall u, v \in V$$

3. Identity element of vector addition:

$$\exists 0, v + 0 = v, \quad \forall v \in V$$

4. Inverse element of vector addition:

$$\forall v, \exists - v, v + (-v) = 0 \quad \forall v \in V$$

5. Scalar Multiplication:

$$a(bv) = (ab)v, \quad \forall v \in V, a, b \in \mathcal{F}$$

6. Identity element of scalar multiplication:

$$\exists 1, 1v = v$$

7. Distributivity of scalar multiplication with respect to vector addition:

$$a(u+v) = au + av, \quad u, v \in V, a \in \mathcal{F}$$

8. Distributivity of scalar multiplication with respect to field addition:

$$(a+b)v = av + bv \quad a, b \in \mathcal{F}, v \in V$$

2.1.1 Orthogonal and Orthonormal Matrices

If A is a matrix, $AI = IA = A$. Where I is the identity matrix.

If $A = [a_1, \ldots, a_m]$ is a matrix where a_i are the columns then;

Definition. 1. A is orthogonal if $a_i \cdot a_j = 0$ if $i \neq j$

2. A is orthonormal if $a_i \cdot a_j = \delta_{ij}$

2.2 Linear Dependence and Linear Independence

Definition. Linear Combination Let V be a vector space over the scalar field $\mathcal{F}$ and let $S = (x_1, \ldots x_n) \subseteq V$ be any n vectors in V. Given a set of scalars $(a_1, \ldots a_n)$, the vector:

$$\sum_{i=1}^{n} a_i x_i \tag{2.1}$$

is called a linear combination of the n vectors $x_i \in S$, with $\mathcal{S}$ of all such linear combinations of elements S is called the subspace spanned by S.

Definition. Linear Dependence Vectors are called linearly dependent if there exists scalars $a_1, \ldots a_n \in \mathcal{F}$, not all equal to 0 such that the linear combination of $x_1, \ldots x_n$:

$$\sum_{i=1}^{n} a_i x_i = 0 \tag{2.2}$$

Definition. Linear Independence Vectors are called linearly independent if it is not linearly dependent

Exercise For $v = (1, 2, 3)$

1. What is the norm of v.

Solution.

$$\begin{aligned}\|v\| &= \sqrt{1+4+9}\\ &= \sqrt{14}\end{aligned}$$

□

2. What is the unit vector in the direction of v:

Solution.

$$\begin{aligned}\hat{v} &= \frac{v}{\|v\|}\\ &= \left(\frac{1}{\sqrt{14}}, \frac{2}{\sqrt{14}}, \frac{3}{\sqrt{14}}\right)\end{aligned}$$

□

3. What are the orthogonal bases for v:

Solution. Use the Cartesian unit vectors: $(1,0,0),(0,1,0),(0,0,1)$ □

4. Find a perpendicular to v.

Solution. Let us choose an arbitrary vector u such that it is linearly independent of v. Without loss of generality:

$$u = (3,5,7) \tag{2.3}$$

Then we take the cross product of these two vectors to find a new vector w which is is perpendicular to both u and v. From this we can see that $w = (-1,2,-1)$ is perpendicular to v through a verification of $v \cdot w = 0$ as required. □

Definition. Cross Product The cross product of two vectors u and v in the basis (i,j,k) is defined to be:

$$u \times v = \begin{vmatrix} i & j & k \\ u_i & u_j & u_k \\ v_i & v_j & v_k \end{vmatrix} \tag{2.4}$$

2.3 Determinant

The determinant of a square matrix A, usually denoted as $\det(A)$ or $|A|$ is a value that is computed from the elements of the matrix. It provides information on the scaling factors on the transformation described by the matrix. Furthermore, it provides us insight into the invertibility of the matrix in that:

$$\det(A^{-1}) = \frac{1}{\det(A)} \tag{2.5}$$

Definition. Determinant For a given matrix A;

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

The determinant $|A|$ is given by;

$$\begin{aligned} |A| &= \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} \\ &= a(ei - hf) - b(di - fg) + c(dh - eg) \end{aligned}$$

2.4 Rank of a matrix

Definition. Given a matrix M, the rank of a matrix is the maximum number of linearly independent columns or rows. (To see this perform Elementary Row Operations (ERO) on the matrix until Row Echelon Form is produced), then the amount of linearly independent columns or rows should be immediately obvious.

Proof that Column Rank is equivalent to Row rank. Let us consider a matrix A such that A is of size (m, n)

Transpose of a matrix

Definition. The transpose of a matrix A, denotd A^T or A' is another matrix B such that:

$$B(i, j) = A(j, i) \tag{2.6}$$

2.5 Decompositions

Eigenvalue Decomposition

For a given square matrix A we say that λ is an eigenvalue and u is an eigenvector if

$$Au = \lambda u, \quad u \neq 0$$

Then for such a given matrix A with n linearly independent eigenvectors $u_1, \dots u_n$. We can factorise A as the following:

$$A = U \Lambda U^{-1}$$

If A is symmetric then its eigenvalues λ_i are real and all its eigenvectors are orthonormal;

$$u_i^T u_j = \delta_{ij}$$

Recall that for orthnormal matrices, the inverse is equal to the transpose and as such:

$$U^TU = UU^T = I$$
$$|U| = 1$$

From which we have that:

$$A = U\Lambda U^T = \sum_{i=1}^{n} \lambda_i u_i u_i^T$$

Given a matrix $X \in \mathbb{R}^{n\times d}$, the principle components of X are the eigenvectors of X^TX. The method of principle component analysis finds eigenvalues and eigenvecotrs of this matrix X^TX

Exercise
Find eigenvalues and eigenvectors for the following matrix:

$$\begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix}$$

Solution. We first find the Characteristic Polynomial:

$$\begin{vmatrix} \lambda-3 & 2 \\ 2 & \lambda-6 \end{vmatrix} = 0$$
$$(\lambda-3)(\lambda-6)-4 = 0$$
$$\lambda^2-9\lambda+14 = 0$$
$$(\lambda-7)(\lambda-2) = 0$$

Therefore the eigenvalues are $\lambda = 2, 7$ We can calculate the eigenvectors through substuiting the eigenvalues into $(A-\lambda I)v = 0$. From this we can see that for $\lambda = 2$ the eigenspace generated by $(1, 2)$. Furthermore for $\lambda = 7$, the eigenspace is spanned by basis $(2, -1)$. (It means the same thing to span an eigenspace and generate an eigenspace) □

2.6 Singular Value Decomposition

Given any real matrix X of size (m, n), it can be expressed as:

$$X = U\Sigma V^T$$

Where U is a $m \times r$ column orthonormal matrix, V^T is a $r \times m$ column-orthonormal matrix and Σ is a diagonal $r \times r$ matrix of singular values λ_i. From this we can see that:

$$X = \lambda_1 u_1 v_1^t + \lambda_2 u_2 v_2^t + \dots \lambda_r u_r v_r^t \tag{2.7}$$

We can consider this singular value decomposition as decomposing a linear mapping X into various components visually through X acting upon a circle to form a skewed ellipse. V^t is a rotation of the circle, for which we have Σ which scales the coordinate axis , effectively scaling the circle into an ellipse. Then finally we have U which rotates the ellipse to the skwewed position.

Compression

We know that the size of X is mn, then our decomposition is:

$$\text{size}(U + \Sigma + V) = mr + r + nr \tag{2.8}$$

As such we can see that the compression ratio is :

$$\begin{aligned} \frac{mr + r + nr}{mn} &= \frac{r(m + 1 + n)}{mn} \\ &\approx \left(\frac{rm}{mn}\right) \\ &= \frac{r}{n} \end{aligned}$$

To achieve better compression, we look at the singular values λ_i and we arrange them in decreasing order:

$$\lambda_i \geq \cdots \geq \lambda_r > 0$$

then for $\lambda_k \gg \lambda_r$

$$\begin{aligned} X &= \sum_{i=1}^{r} \lambda_i u_i v_i^t \\ &\approx \sum_{i=1}^{k} \lambda_i u_i v_i^t \\ &= \hat{X} \end{aligned}$$

for which we have compression ratio k/n

Chapter 3

Basics of Probability Theory and Baye's Rule

3.1 Probability Theory

We ask ourselves the question as to the motivation of proability theory. Laplace a French Mathematician states that ”Probability is common sense” reduced to calculation”. we use probability theory is useful in understanding. studying, and analysing complex real world systems.

3.1.1 Understanding Uncertainty

We have several ways to describe probabilities and associated uncertainties.

1. Aleatory: Random, left to choice with no real ability to predict outcome.
2. Epistemic: Encoding Knowledge as a measure of belief, ability to predict outcome
3. Sensing: ability to encode noisy measurements.

It should be noted that it is better to be imprecisely right than precisely wrong!

3.1.2 How good is the output h_s

As data scientists it is important to consider the following questions.

1. We learn our hypothesis from given training data. How good is it for the test data?
2. Given a previously unseen data, how will h_s perform on it?

Predictions and Probabilities

When we make predictions we should assign probabilities with the predicitions. For example we can say that:

1. There is 20% chance it will rain tomorrow.
2. There is 50% chance that the tumour is malignant.
3. There is 60% chance that the stock market will fall by the end of the week.

4. There is 30% that the next president of the United states will be a Deomcrat.

We can assign probabilities to complex and complicated events using the correct data algorithms and through counting.

3.2 Probability Basics

Experiments and Sample Space

Consider an experiment and let S be the space of possible outcomes.

For example we can say that for:

1. Experiment tossing a coin; $S = \{h, t\}$
2. Experiment rolling a pair of dices; $S = \{(1,1), (1,2), \ldots, (6,6)\}$
3. Experiment is a race consisting of three cars with labels: $\{(1,2,3), (1,3,2), \ldots (3,2,1)\}$

Let us take another example: $S = \{1, 2, \ldots, m\}$. Where for non negative probabilities $p_i \geq 0$ for $i = 1, \ldots, m$ the sum of probabilities is unitary; $\sum_i p_i = 1$, with p_i the probability of outcome i. Lets say that we toss a fair coin, then we have as above that the sample space is $S = \{h, t\}$. Then we have that $p_h = 0.5$ and $p_t = 0.5$

Events

> **Definition.** An Event A is a set of possible outcomes of the experiment. Therefore we have $A \subseteq S$.

We offer the example of A being the event of getting a seven when we roll a pair of dice. Then we have that;

$$A = \{(1,6), (6,1), (2,5), (5,2), (4,3), (3,4)\}$$

Hence we have the probability of A is;

$$\begin{aligned} P(A) &= \frac{6}{6^2} \\ &= 6 \end{aligned}$$

In general we have that:

$$P(A) = \sum_{i \in A} p_i \tag{3.1}$$

Events and Sample Space

Both the sample space S and events A are probability sets. We define the follwing:

> **Definition.**
>
> $$P(S) = 1; P(\phi) = 0 \tag{3.2}$$

Where the proability of all the entire sample space is bound to be 1 and the probability of no events ϕ (the empty set) is then 0. In addition we define the union of the probability of events as follow:

$$P(A \cup B)f = P(A) + P(B) - P(A \cap B) \tag{3.3}$$

$$\tag{3.4}$$

Often we denote the intersection in differing notations:

$$P(A \cap B) \equiv P\left(AB\right) \equiv P(A, B) \tag{3.5}$$

Definition. We define the complement of an event A as:

$$P(A^c) = 1 - P(A) \tag{3.6}$$

The above denotes what is called the axioms of probability.

Let us denote this with an example:

Example: Suppose the probability of raining today is 0.4 and tomorrow is also 0.4 and on both days is 0.1. What is the probability it does not rain on either day?

Let us denote the sample space S as $S = \{(R, N)(R, R), (N, N)(N, r)\}$. Let A be the event that it will rain today and B that it will rain tomorrow. Then

$$A = \left\{(R, N), (R, R)\right\}, \left\{(N, N), (N, R)\right\} \tag{3.7}$$

Then the probability that it will rain at least today or tomorrow is:

$$P(A \cup B) = 0.4 + 0.4 - 0.1 = 0.7 \tag{3.8}$$

Therefore the probability that it will not rain on either day is 0.3.

3.2.1 Discrete Random Variables

Events like "ASX is up" are binary events. We can extend the definition of such events by defining a discrete random variable. We can then say that the probability $P(X = x)$ is the probability that event $X = x$. For a discrete random variable,

$$0 \leq P(X = x) \leq 1$$
$$\sum_{x \in X} P(X = x) = 1$$

3.2.2 Continuous Random Variables

Random Variables can also be continuous, examples of this are; "Height, rainfall, salary, chemical conccentration etc . . . . ". we can talk about the average (mean) and standard deviation or variance.

3.2.3 Probability Densities

Random Variables (both continous and discrete) are associated with distributions.

1. Common examples of discrete distributions are Bernoulli, binomial, multinomial, Poisson.

2. Common examples of continuous distributions are Gaussian, Laplacian, Exponential, Gamma.

Associated with distributions are parameters. In Machine Learning and Statistics, we seek to learn the parameters of a distribution from data. We can talk about the probability distribution function:

$$p(x \in (a, b)) = \int_a^b p(x)dx$$

For which the cumulative distriubtion function

$$P(z) = \int_{-\infty}^{z} p(x)dx$$

and for $p(x) \geq 0$

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

Expectations

The expectation of a discrete random variable is given by;

$$E[f] = \sum_x p(x)f(x) \tag{3.9}$$

The expectation of a continous random variable is given by;

$$E[f] = \int p(x)f(x)dx \tag{3.10}$$

The conditional expectation of a discrete random variable is given by;

$$E_x[f|y] = \sum_x p(x|y)f(x) \tag{3.11}$$

The approximate expectation of both discrete and continuous random variables is given by;

$$E[f] \approx \frac{1}{N}\sum_{n=1}^{N} f(x_n) \tag{3.12}$$

Variance and Covariance

The variance of a random variable f is given by:

$$\mathrm{Var}[f] = E\left[\left(f(x) - E\left[f(x)\right]\right)^2\right] = E\left[f(x)^2\right] - E\left[f(x)\right]^2 \tag{3.13}$$

The covariance of two random scalar variables x and y is given by;

$$\begin{aligned}\mathrm{cov}[x,y] &= E_{x,y}\left[\{x-E[x]\}\{y-E[y]\}\right]\\ &= E_{x,y}[xy]-E[x]E[y]\end{aligned}$$

The covariance of two random vector variables $\vec{x}$ and $\vec{y}$ is given by;

$$\begin{aligned}\mathrm{cov}[\vec{x},\vec{y}] &= E_{\vec{x},\vec{y}}\left[\{\vec{x}-E[\vec{x}]\}\left\{\vec{y}^T-E[\vec{y}^T]\right\}\right]\\ &= E_{\vec{x},\vec{y}}[\vec{x}\vec{y}^T]-E[\vec{x}]E[\vec{y}^T]\end{aligned}$$

INCOMPLETE —– Will complete given time