CHAPTER 2

Mathematical Review

Prior to discussing machine learning, a brief overview of statistics is necessary. Broadly, *statistics* is the analysis and collection of quantitative data with the ultimate goal of making actionable insights on this data. With that being said, although machine learning and statistics aren't the same field, they are closely related. This chapter gives a brief overview of terms relevant to our discussions later in the book.

Statistical Concepts

No discussion about statistics or machine learning would be appropriate without initially discussing the concept of probability.

Probability

Probability is the measure of the likelihood of an event. Although many machine learning models tend to be deterministic (based off of algorithmic rules) rather than probabilistic, the concept of probability is referenced specifically in algorithms such as the expectation maximization algorithm in addition to more complex deep learning architectures such a recurrent neural networks and convolutional neural networks. Mathematically, this algorithm is defined as the following:

$$Probability\ of\ Event\ A = \frac{number\ of\ times\ event\ A\ occurs}{all\ possible\ events}$$

This method of calculating probability represents the *frequentist* view of probability, in which probability is by and large derived from the following formula. However, the other school of probability, Bayesian, takes a differing approach. Bayesian probability theory is based on the assumption that probability is conditional. In other words, the likelihood of an event is influenced by the conditions that currently exist or events that have happened prior. We define conditional probability in the following equation. The probability of an event A, given that an event B has occurred, is equal to the following:

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

T. Beysolow II, *Introduction to Deep Learning Using R*, DOI 10.1007/978-1-4842-2734-3_2

$$\text{Provided } P(B) > 0.$$

In this equation, we read $P(A|B)$ as "the probability of A given B" and $P(A \cap B)$ as "the probability of A and B."

With this being said, calculating probability is not as simple as it might seem, in that dependency versus independency must often be evaluated. As a simple example, let's say we are evaluating the probability of two events, A and B. Let's also assume that the probability of event B occurring is dependent on A occurring. Therefore, the probability of B occurring should A not occur is 0. Mathematically, we define dependency versus independency of two events A and B as the following:

$$P(A|B) = P(A)$$

$$P(B|A) = P(B)$$

$$P(A \cap B) = P(A)P(B)$$

In Figure 2-1, we can envision events A and B as two sets, with the union of A and B as the intersection of the circles:

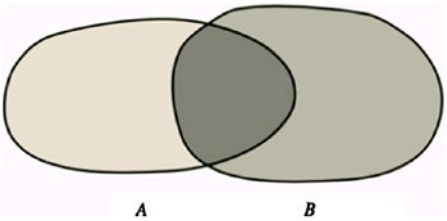

Figure 2-1. *Representation of two events (A,B)*

Should this equation not hold in a given circumstance, the events A and B are said to be dependent.

And vs. Or

Typically when speaking about probability—for instance, when evaluating two events A and B—probability is often in discussed in the context of "the probability of A *and* B" or "the probability of A *or* B." Intuitively, we define these probabilities as being two different events and therefore their mathematical derivations are difference. Simply stated, *or* denotes the addition of probabilities events, whereas *and* implies the multiplication of probabilities of event. The following are the equations needed:

And (multiplicative law of probability) is the probability of the intersection of two events A and B:

$$P(A \cap B) = P(A)P(B|A)$$

$$= P(B)P(A|B)$$

If the events are independent, then

$$P(A \cap B) = P(A)P(B)$$

Or (additive law of probability) is the probability of the union of two events A and B:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

The symbol $P(A \cup B)$ means "the probability of A or B."

Figure 2-2 illustrates this.

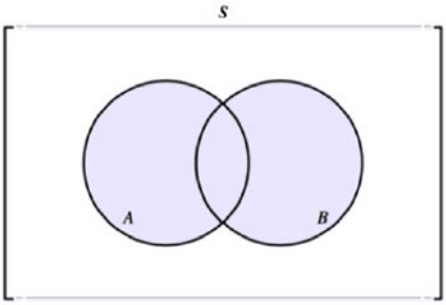

Figure 2-2. *Representation of events A,B and set S*

The probabilities of A *and* B exclusively are the section of their respective spheres which do not intersect, whereas the probability of A *or* B would be the addition of these two sections plus the intersection. We define S as the sum of all sets that we would consider in a given problem plus the space outside of these sets. The probability of S is therefore always 1.

With this being said, the space outside of A and B represents the opposite of these events. For example, say that A and B represent the probabilities of a mother coming home at 5 p.m. and a father coming home at 5 p.m. respectively. The white space represents the probability that neither of them comes home at 5 p.m.

Bayes' Theorem

As mentioned, Bayesian statistics is continually gaining appreciation within the fields of machine learning and deep learning. Although these techniques can often require considerable amounts of hard coding, their power comes from the relatively simple theoretical underpinning while being powerful and applicable in a variety of contexts. Built upon the concept of conditional probability, Bayes' theorem is the concept that the probability of an event A is related to the probability of other similar events:

$$P\left(B_j \mid A\right) = \frac{P\left(A \mid B_j\right)P\left(B_j\right)}{\sum_i^k P\left(A \mid B_i\right)P\left(B_i\right)}$$

Referenced in later chapters, Bayesian classifiers are built upon this formula as well as the expectation maximization algorithm.

Random Variables

Typically, when analyzing the probabilities of events, we do so within a set of random variables. We define a random variable as a quantity whose value depends on a set of possible random events, each with an associated probability. Its value is known prior to it being drawn, but it also can be defined as a function that maps from a probability space. Typically, we draw these random variables via a method know as random sampling. *Random sampling* from a population is said to be random when each observation is chosen in such a way that it is just as likely to be selected as the other observations within the population.

Broadly speaking, the reader can expect to encounter two types of random variables: *discrete random variables* and *continuous random variables*. The former refers to variables that can only assume a finite number of distinct values, whereas the latter are variables that have an infinite number of possible variables. An example is the number of cars in a garage versus the theoretical change in percentage change of a stock price. When analyzing these random variables, we typically rely on a variety of statistics that readers can expect to see frequently. But these statistics often are used directly in the algorithms either during the various steps or in the process of evaluating a given machine learning or deep learning model.

As an example, arithmetic means are directly used in algorithms such as K-means clustering while also being a theoretical underpinning of the model evaluation statistics such as mean squared error (referenced later in this chapter). Intuitively, we define the arithmetic mean as the central tendency of a discrete set of numbers—specifically it is the sum of the values divided by the number of the values. Mathematically, this equation is given by the following:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

The *arithmetic* mean, broadly speaking, represents the most likely value from a set of values within a random variable. However, this isn't the only type of mean we can use to understand a random variable. The *geometric* mean is also a statistic that describes the central tendency of a sequence of numbers, but it is acquired by using the product of the values rather than the sum. This is typically used when comparing different items within a sequence, particularly if they have multiple properties individually. The equation for the geometric mean is given as follows:

$$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \left(x_1 * x_2 * \ldots * x_n\right)^{\frac{1}{n}}$$

For those involved in fields where the use of time series is frequent, geometric means are useful to acquiring a measure of change over certain intervals (hours, months, years, and so on). That said, the central tendency of a random variable is not the only useful statistic for describing data. Often, we would like to analyze the degree to which the data is dispersed around the most probable value. Logically, this leads us to the discussion of variance and standard deviation. Both of these statistics are highly related, but they have a few key distinctions: *variance* is the squared value of standard deviation, and the standard deviation is often more referenced than variance across various fields. When addressing the latter distinction, this is because variance is much harder to visually describe, in addition to the fact that the units that variance is in are ambiguous. Standard deviation is in the units of the random variable being analyzed and is easy to visualize.

For example, when evaluating the efficiency of a given machine learning algorithm, we could draw the mean squared error from several epochs. It might be helpful to collect sample statistics of these variables, such that we can understand the dispersion of this statistic. Mathematically, we define variance and standard deviation as the following

Variance

$$\sigma^2 = \frac{\Sigma\left(X-\mu\right)^2}{N}$$

$$Var(X) = E\left[\left(X - E\left(\left[X\right]\right)\right)^2\right]$$

$$= E[X^2] - 2XE\left[X\right] + \left(E\left[X\right]\right)^2$$

$$= E\left[X^2\right] - 2E\left[X\right]E\left[X\right] + \left(E\left[X\right]\right)^2$$

$$= E\left[X^2\right] - 2E\left[X\right]E\left[X\right] + \left(E\left[X\right]\right)^2$$

Standard Deviation

$$\sigma = \sqrt{\left(\frac{\sum_{i}^{n} (x_i - \bar{x})^2}{n-1} \right)}$$

Also, covariance is useful for measuring the degree to which a change in one feature affects the other. Mathematically, we define covariance as the following:

$$cov(X,Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Although deep learning has made significant progress in modeling relationships between variables with non-linear correlations, some estimators that one would use for more simple tasks require this as a preliminary assumption. For example, linear regression requires this to be an assumption, and although many machine learning algorithms can model complex data, some are better at it than others. As such, it is recommended that prior to selecting estimators features be examined for their relationship to one another using these prior statistics. As such, this leads us to the discussion of the correlation coefficient which measures the degree to which to variables are linearly related to each other. Mathematically, we define this as follows:

$$correlation = \rho = \frac{1}{n} \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

Correlation coefficients can have a value as low as –1 and as high as 1, with the lower bound representing an *opposite* correlation and the upper bound representing *complete* correlation. A correlation coefficient of 0 represents complete lack of correlation, statistically speaking. When evaluating machine learning models, specifically those that perform regression, we typically reference the coefficient of determination (R squared) and mean squared error (MSE). We think of *R squared* as a measure of how well the estimated regression line of the model fits the distribution of the data. As such, we can state that this statistic is best known as the *degree of fitness* of a given model. MSE measures the average of the squared error of the deviations from the models predictions to the observed data. We define both respectively as the following:

Coefficient of Determination (R Squared)

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

Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$$

With respect to what these values should be, I discuss that in detail later in the text. Briefly stated, though, we typically seek to have models that have high R squared values and lower MSE values than other estimators chosen.

Linear Algebra

Concepts of linear algebra are utilized heavily in machine learning, data science, and computer science. Though this is not intended to be an exhaustive review, it is appropriate for all readers to be familiar with the following concepts at a minimum.

Scalars and Vectors

A *scalar* is a value that only has one attribute: *magnitude*. A collection of scalars, known as a vector, can have both magnitude and direction. If we have more than one scalar in a given vector, we call this an *element of vector space*. Vector space is distinguished by the fact that it is sequence of scalars that can be added and multiplied, and that can have other numerical operations performed on them. *Vectors* are defined as a column vector of n numbers. When we refer to the indexing of a vector, we will describe *i* as the index value. For example, if we have a vector x, then x_1 refers to the first value in vector x. Intuitively, imagine a vector as an object similar to a file within a file cabinet. The values within this vector are the individual sheets of paper, and the vector itself is the folder that holds all these values.

Vectors are one of the primary building blocks of many of the concepts discussed in this text (see Figure 2-3). For example, in deep learning models such as Doc2Vec and Word2Vec, we typically represent words, and documents of text, as vectors. This representation allows us to condense massive amount of data into a format easy to input to neural networks to perform calculations on. From this massive reduction of dimensionality, we can determine the degree of similarity, or dissimilarity, from one document to another, or we can gain better understanding of synonyms than from simple Bayesian inference. For data that is already numeric, vectors provide an easy method of "storing" this data to be inputted into algorithms for the same purpose. The properties of vectors (and matrices), particularly with respect to mathematical operations, allow for relatively quick calculations to be performed over massive amounts of data, also presenting a computational advantage of manually operating on each individual value within a data set.

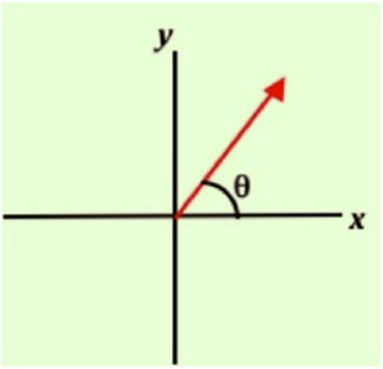

Figure 2-3. *Representation of a vector*

Properties of Vectors

Vector dimensions are often denoted by $\mathbb{R}^n$ or $\mathbb{R}^m$ where n and m is the number of values within a given vector. For example, $x \in \mathbb{R}^5$ denotes set of 5 vectors with real components. Although I have only discussed a column vector so far, we can also have a row vector. A transformation to change a column vector into a row vector can also be performed, known as a transposition. A *transposition* is a transformation of a matrix/vector X such that the rows of X are written as the columns of X^T and the columns of X are written as the rows of X^T.

Addition

Let's define two vectors $d = [d_1, d_2, \ldots, d_n]^T$ and $e = [e_1, e_2, \ldots, e_n]^T$ where

$$d_n = e_n, \text{ for } i = 1, 2, \ldots, n$$

The sum of the vectors is therefore the following:

$$d + e = \left[(d_1 + e_1), (e_2 + d_2), \ldots, (d_n + e_n)\right]^T$$

Subtraction

Given that the assumptions from the previous example have not changed, the difference between vectors d and e would be the following:

$$d - e = \left[(d_1 - e_1), (e_2 - d_2), \ldots, (d_n - e_n)\right]^T$$

Element Wise Multiplication

Given that the assumptions from the previous example have not changed, the product of vectors d and e would be the following:

$$d * e = \left[(d_1 * e_1), (e_2 * d_2), \ldots, (d_n * e_n) \right]^T$$

Axioms

Let *a*,*b*, and *x* be a set of vectors within set *A*, and *e* and *d* be scalars in B. The following axioms must hold if something is to be a vector space:

Associative Property

The associative property refers to the fact that rearranging the parentheses in a given expression will not change the final value:

$$x + (a + b) = (x + a) + b$$

Commutative Property

The commutative property refers to the fact that changing the order of the operands in a given expression will not change the final value:

$$a + b = b + a$$

Identity Element of Addition

$$a + 0 = a, \textit{ for all } a \in A$$

Where $0 \in A$. 0 in this instance is the zero vector, or a vectors of zeros.

Inverse Elements of Addition

In this instance, for every a := A, there exists an element –a := A, which we label as the additive inverse of a:

$$a + (-a) = 0$$

Identity Element of Scalar Multiplication

$$(1)a = a$$

Distributivity of Scalar Multiplication with Respect to Vector Addition

$$e(a+b) = ea + eb$$

Distributivity of Scalar Multiplication with Respect to Field Addition

$$(a+b)d = ad + bd$$

Subspaces

A *subspace* of a vector space is a nonempty subset that satisfies the requirements for a vector space, specifically that linear combinations stay in the subspace. This subset is "closed" under addition and scalar multiplication. Most notably, the zero vector will belong to every subspace. For example, the space that lies between the hyperplanes of produced by a support vector regression, a machine learning algorithm I address later, is an example of a subspace. In this subspace are acceptable values for the response variable.

Matrices

A matrix is another fundamental concept of linear algebra in our mathematical review. Simply put, a *matrix* is a rectangular array of numbers, symbols, or expressions arranged in rows and columns. Matrices have a variety of uses, but specifically are often used to store numerical data. For example, when performing image recognition with a convolutional neural network, we represent the pixels in the photos as numbers within a 3-dimensional matrix, representing the matrix for the red, green, and blue photos comprised of a color photo. Typically, we take an individual pixel to have 256 individual values, and from this mathematical interpretation an otherwise difficult-to-understand representation of data becomes possible. In relation to vectors and scalars, a matrix contains scalars for each individual value and is made up of row and column vectors. When we are indexing a given matrix A, we will be using the notation A_{ij}. We also say that $A = a_{ij,}\ A \in \mathbb{R}^{m \times n}$.