Unit 1.1

Basics linear algebra, probability, calculus

IST 718 – Big Data Analytics

Daniel E. Acuna

http://acuna.io

Scalars

- Represented by Greek letters α, β, γ
- Represent numbers
- $\alpha = 0.1, \beta = 1^{-10}$

Notation and simple matrix algebra

- We let $\mathbf{X}$ denote a $n \times p$ matrix whose (i, j)th element is x_{ij}. That is,

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

- $\mathbf{X}$ can be visualize as a spreadsheet of numbers with n rows and p columns.

- The rows of $\mathbf{X}$ can be written as $x_1, x_2, \ldots, x_n$. Here x_i is a vector of lenght p, containing the p variable measurements for the ith observation. That is,

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

- IMPORTANT: vectors are by default represented as columns.

- The columns of $\mathbf{X}$ can written as $x_1, x_2, \ldots, x_p$. Each is a vector of length n. That is,

$$\mathbf{x}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}$$

Using the previous notation, the matrix $\mathbf{X}$ can be written as

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_p \end{pmatrix} \qquad \text{or} \qquad \mathbf{X} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}$$

The ^{T} notation denotes the *transpose* of a matrix or vector. So, for example,

$$\mathbf{X}^T = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{bmatrix}$$

while

$$x_i^T = \begin{pmatrix} x_{i1} & x_{i2} & \cdots & x_{ip} \end{pmatrix}$$

- We use y_i to denote the ith observation of the variable on wish we wish to make predictions. Hence we write the set of all n observations in *vector form* as

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

- Then our observed data consist of $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where each x_i is a vector of length p.
- If $p = 1$, then x_i is simply a scalar.

- In this course, a vector of length n will always be denoted in *lower case bold;* e.g.

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$$

- However, vectors that are not of length n (e.g., x_i) will be denoted in *lower case*. The same rule applies to scalars (e.g., a).
- Matrices will be denoted using *bold capitals*, such as $\mathbf{X}$
- Random variables will be denoted using capitals, e.g. A

Matrix

- Sometimes, we can define a matrix by its components as follows $\mathbf{A} = (f(i,j))_{ij}$ where $f(i,j)$ is a function of i and j.
- For example, define the matrix

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 0 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

using a function

Matrix operations

- Scalar times matrix: $\alpha\mathbf{A} = (\alpha \times a_{ij})_{ij}$
- Matrix addition: $\mathbf{A} + \mathbf{B}$ (add each element one at a time)
- Matrix multiplication: $\mathbf{AB}$ ($\#\mathrm{cols}_A = \#\mathrm{rows}_B$)
$$\mathbf{AB} = \left(\sum_z a_{iz}b_{zj}\right)_{ij}$$
- Matrix transposition: make rows the columns
$$\mathbf{A}^T = (a_{ij})_{ji}$$
- Many operations can be easily written as matrices

Special matrices and properties

- Identity matrix (diagonal values are 1, everything else is 0)
- $I = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}$
- Matrix inverse: $AA^{-1} = I$
- Matrix addition is commutative: $A + B = B + A$
- Matrix multiplication is NOT commutative: $AB \neq BA$
- $(AB)^T = B^T A^T$
- Other matrix properties https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf (https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf)

Dimension

To indicate that an object is:

- a scalar, we will use the notation $a \in \mathbb{R}$
- a vector of length n, we will use $\mathbf{a} \in \mathbb{R}^{\mathbb{n}}$
- a vector of length k, we will use $a \in \mathbb{R}^{\mathbb{k}}$
- a $r \times s$ matrix, we will use $\mathbf{A} \in \mathbb{R}^{r \times s}$

Interpreting graphs (1)

- Equation of the line: intercept, slope
  - Interpreting intercept and slope
- Application:
  - Model 1: $\widehat{income} = f(age) = 20000 + 5000 \times age$
  - The unit of the intercept is different from the unit of the slope
  - Using matrix notation to make predictions for $age = \{20, 25, 40\}$
    - Represent model as a vector $b = \begin{pmatrix} 20000 \\ 5000 \end{pmatrix}$
    - Represent data as matrix $X = ?$
    - Making predictions: $X \times b$

Interpreting graphs (2)

- Model 2:
  $\widehat{income} = f(age) = 20000 + 5000 \times age + 10000 \times education$
- Represent model 2 as a matrix?

Optimization for model fitting

- Let's assume a simple model where we are trying to predict age
$$\widehat{age} = f() = b$$
- This model does not take any features or inputs
- We would like to find the b to predict well the following $ages = \{20, 25, 40\}$
- We would like to quadratically penalize mistakes, i.e.: $(\widehat{age} - age)^2$
- How do we find the right parameters for the model?

Optimization

- We can find a minimum or maximum of a function by looking at the slope
- Finding the minimum of a function:

$$\frac{df(x)}{dx} = 0$$

- In multiple dimensions it is called a gradient:

$$g = \left(\frac{df(x_1)}{dx_1} \; \frac{df(x_2)}{dx_2} \; \cdots \; \frac{df(x_p)}{dx_p} \right)^T$$

Derivatives

- Definition of the derivative:
$$\frac{df(x)}{d(x)} \approx \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}$$
- This means: the infinitesimal change in the function as the change is taken to zero
- Take as an examples:
  - $f_1(x) = a + xb$
  - $f_2(x) = x^2$

Optimization for model fitting

- Model: $\widehat{age} = b$
- Data: $ages = \{20, 25, 40\}$
- Function to minimize with quadratic errors?
- Optimal value for b?

Optimization of a general model

- Model: $y = b_0 + \sum b_i x_i$
- Data: set of features X and outputs or targets y
- Function to minimize?
- Optimal value for b?

Other common derivation rules

- Chain rule:
$$\frac{dg(f(x))}{dx} = \frac{dg(f)}{f}\frac{df(x)}{x}$$

- Exercise, combine the following rules:

(1) $\frac{d(cf(x))}{dx} = c\frac{df(x)}{dx}$

(2) $\frac{d(f(x)+g(x))}{dx} = \frac{d(f(x))}{dx} + \frac{g(g(x))}{dx}$

(3) $\frac{d(x^n)}{dx} = nx^{n-1}$

to solve $\frac{d(5x-\mu)^3}{dx}$

Common properties

- $\dfrac{d(e^x)}{dx} = e^x$
- $\dfrac{d(\log(x))}{dx} = \dfrac{1}{x}$
- A common prediction function for probability values is the sigmoid:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

- Use the properties learned before to compute $\dfrac{d(\sigma(z))}{z}$

A more complicated loss function

- Logistic regression has a loss function called *cross-entropy*:

$$l(z) = -y\log(\sigma(z)) - (1-y)\log(1-\sigma(z))$$

- Compute $\dfrac{dl(z)}{dz}$

Probability

- There are some phenomena which are not certain and therefore need a set of tools to still work with them
- Probability deals with the likelihood or chance that an event will occur, and can deal with these phenomena
- Several interpretation of what a likelihood is
  - Frequency: the relative frequency of how many times an event occurs if the same conditions are repeated many times. E.g., if I flip a coin 5M times, what is the relative frequency of heads?
  - Subjective definition: subjects have beliefs about the probability of an outcome which must be updated with certain consistent rules. E.g.: I may assume I will see heads 50% of the time, but I will update my belief if I see tails 10 times in a row

Experiments and events

- An experiment is any process for which the outcome is unknown
- E.g.,:
  - Experiment to estimate, out of 10 coin tosses, the number of times heads will be obtained
  - If a spark job is running on 100 computers, estimate the probability that the job will finish successfully if all computers must finish without error
  - If I estimate that the average age of my data is 30 years, how likely is it to see someone 60 years old in the future?

Set theory

- The collection of all possible outcomes in an experiment is called *sample space*
- For example:
  - The *sample space* of an experiment with a dice could be $S = \{1, 2, 3, 4, 5, 6\}$
  - The event A that an even number is obtained is defined by $A = \{2, 4, 6\}$
- Operations of set theory:
  - Union
  - Intersection
  - Complement

Probability

- Axioms:
  1. Probability of any event is greater or equal to zero
     $p(A) \geq 0$
  2. If an event S is certain to occur, then
     $p(S) = 1$
  3. The probability of an infinite number of independent events $A, B, C, \ldots$
     is the sum of the probability of each event
     $p(A) + p(B) + p(C) + \ldots$

- Any function that follows Axioms 1, 2, and 3 is a *probability distribution*

Some derived properties (1)

- For event A,
  $p(\neg A) = 1 - p(A)$, proof?
  $p(\neg A) = 1 - p(A)$, proof?

- For any two events A and B,
  $p(A \cup B) = p(A) + p(B) - p(A \cap B)$

- Conditional probability (probability of an event knowing that another event is certain)
  $p(A \mid B) = P(A \cap B)/P(B)$

Some derived properties (2)

- Example:
  1. A ball is selected from an urn with red, blue, and green balls. If the probability of red is $^1/_5$ and blue is $^2/_5$, what is the probability of getting a green ball?
  2. A friend tells you that she has two children, then you see one of her children and it is a girl. What is the probability that the other child is **also** a girl?

Random variables and probability distributions

- A random variable is a real-valued function that is defined on a sample space of an experiment
- For example, a function defined over the number of heads after 5 tosses is a random variable. For sample $s = HHTTH$, the random variable would be $X(s) = 3$
- The distribution of a random variable X is the probability of the events underlying the random variable

Discrete and continuous random variables

- If the random variable X can take on a finite number of k different values $x_1, \ldots x_k$ or, an infinite sequence of them, X is a *discrete random variable*
- Random variables that can take on every value on an interval are *continuous random variables*

Discrete probability distribution

- Describes the probability of each real value x of a discrete random variable
$$p(X = x)$$
sometimes denoted simply as $p(x)$
- The set of points such that $\{x \mid p(x) > 0\}$ is denoted the *support* the probability distribution
- The sum of all events must sum up to 1: $\sum_x p(x) = 1$
- p is also call a *probability mass function*

Example of discrete probability distributions

- Bernoulli distribution (probability of tossing head)
$$p(X = H) = p(H) = \theta$$
and since the probability of all events must sum up to one $p(H) + p(T) = 1$ then
$$p(T) = 1 - \theta$$
- This can be compactly represented as
$$p(x) = \theta^x (1 - \theta)^{1-x}$$
if we consider heads as 1 and tails as 0.

Example of discrete probability distributions (2)

- Uniform distribution between integers a and b would be

$$p(x) = \begin{cases} \frac{1}{b-a+1} & a \leq x \leq b \\ 0 & \text{o.w.} \end{cases}$$

Continous probability distribution

- Defines probabilities for bounded closed intervals $[a, b]$

$$p(a \leq X \leq b) = \int_a^b p(x)dx$$

- $p(x) \geq 0$ for all x
- $\int_{-\infty}^{\infty} p(x) = 1$
- A single point in a continuous distribution has probability 0
- p is called a *probability density function*

Example of a continous distribution

- Uniform distribution on an interval

$$p(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{o.w.} \end{cases}$$

Example of a continous distribution (2)

- Sometimes we define probabilities without worrying about whether they sum up to 1

$$p(x) \propto \begin{cases} 4x & 0 \leq x \leq 1 \\ 0 & \text{o.w.} \end{cases}$$

- How to properly define the previous probability distribution? Hint: Use the fact that $\int p(x) = 1$

Example of continous distribution (3)

- Gaussian distribution
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$
- μ is called the mean and σ is called the standard deviation.

Joint distributions

- The joint distribution of a set of random variables
  $$p(X_1 \in C_1, X_2 \in C_2, \ldots, X_k \in C_k)$$
  can be read as the probability that the random variables are simultaneously in the intervals $C_1, \ldots, C_k$

Marginal probability

- From a simple example distribution
$$p(X_1 \in C_1, X_2 \in C_2)$$
we can obtain the following
$$p(X_1 \in C_1) = \sum_{x_2 \in C_2} p(X_1 = x_1, X_2 = x_2)$$
for a discrete distribution, and
$$p(X_1 \in C_1) = \int_{x_2 \in C_2} p(X_1 = x_1, X_2 = x_2) dx_2$$
for a continous distribution.
- This can be generalized for many variables

Conditional probability

- If we did not have uncertainty about the value of random variable X_2, we write
$$p(X_1 \in C_1 \mid X_2 \in C_2) = \frac{p(X_1 \in C_1, X_2 \in C_2)}{p(X_2 \in C_2)}$$

Independence

- If two random events are independent (they don't depend on each other), their joint probability can be expressed as the factor of their distributions

$$p(X_1 \in C_1, X_2 \in C_2) = p(X_1 \in C_1)p(X_2 \in C_2)$$

Common statistics

- Expectation: A *fancy average*

$$E[f(x)] = \sum_x p(x) f(x) \quad E[f(x)] = \int_x p(x) f(x) dx$$

- Variance: *Spread*

$$Var[f(x)] = E[(f(x) - E[f(x)])^2]$$

- Covariance: *Co-spread*

$$Cov(f(x), g(y)) = E[(f(x) - E[f(x)])(g(y) - E[g(y)])]$$

Some exercises

- A friend tells you that she has two children. You see one children and it is a girl.
  - What is the probability that the other child is **also** a girl?
  - What is the probability that the other child is a girl?
- You toss 2 die
  - What is the probability that sum of the die is 2?
  - If I pay you one dollar for each and 1 or 6. What is expected value you are expected to receive?