Introduction to Data Science

Math and Probability Prerequisites

Collin Cademartori

October 9, 2020

Plan for This Lecture

- Review topics from high school math
- Overview of probability concepts
  - Arithmetic of probability
  - Bayes rule
  - Some common distributions
- Will try to make these topics interesting
- Please stop me if I go over anything too quickly

Zoom Lecture Logistics

- Please turn on your cameras if you can
- I will periodically ask the class questions
- Please use the 'raise hand' function in Zoom to answer
- The 'raise hand' button can be found at the bottom of the participants list
- If you have a question, please just interrupt me
- I'll try to watch the chat, but speaking is better if you can

Pre-Calculus in Three Slides

- We will encounter many the common functions in this class
- **Polynomials:** $p(x) = a_k x^k + a_{k-1} x^{k-1} + \cdots + a_1 x + a_0$
  - The highest power in the polynomial is its **degree**
  - If $p(x_0) = 0$, then x_0 is called a **root** or zero of p
  - The **multlicity** of a root x_0 is the largest j so that

$$p(x) = (x - x_0)^j p_1(x)$$

    for some polynomial p_1
  - **Fundamental theorem of algebra:**

    A degree k polynomial has k (possibly complex) roots (counted with multiplicity)
  - A polynomial is *determined up to a constant* by its roots
  - Statistically, polynomials are good at modeling functions:
    - That don't vary too rapidy
    - That don't take values across many orders of magnitude

Pre-Calculus in Three Slides

- **Trigonometric Functions:** $\sin(x)$ and $\cos(x)$
  - These functions are periodic waves
  - We can control the properties of this wave

$$f(x) = c + A\sin(\omega x + \psi)$$

    - Changing c shifts the function up and down
    - Changing ψ shift the function left and right
    - A controls the **amplitude** of the wave
    - ω controls the **frequency** of the wave
  - Sums of these represent more complex periodic functions
  - Arbitrary functions can be well approximated by such sums
    - This is a result from the field of Fourier analysis
  - Statistically, trig functions can model periodic phenomena
    - Ex: time series with seasonal trends

Pre-Calculus in Three Slides

- **Exponential Functions:** a^x
  - Exponential functions grow faster than any polynomial
  - If $a = e$, we refer to this as 'the' exponential function $\exp(x)$
  - $\exp(x)$ has series representation $\sum_{n=0}^{\infty} \frac{x^n}{n!}$
  - Exponential functions can model compound growth
    - Ex: the spread of a virus after initial introduction
- **Logarithmic Functions:** $\log_a(x)$
  - $\log_a(x)$ is the number b such that $a^b = x$ (defined for $x > 0$)
  - In other words, $\log_a(x)$ is inverse of a^x
  - When $a = e$, this is the natural logarithm $\log(x)$

Calculus Review

- We will make frequent use of basic calculus in statistics
- We will review these concepts using convergence rates
- Suppose that $f(x)$ and $g(x)$ are two functions and that

$$\lim_{x \to 0} \frac{f(x)}{g(x)} = 0$$

  Then we write $f(x) = o(g(x))$ and say f is "little o" of g
- If $f(x) \to 0$ and $g(x) \to c \neq 0$ as $x \to 0$, then $f = o(g)$
- If $f(x), g(x) \to 0$, we may or may not get $f = o(g)$
- In this case, if $f = o(g)$, then f goes to zero faster than g
- This is a statement about the rate at which f goes to zero
- We can restate calculus results in terms of these rates

Derivative as The Local Linear Part of $f(x)$

- We can characterize the derivative of f at a as follows
- $f'(a)$ is the unique value such that

$$\lim_{h \to 0} \frac{|f(a+h) - (f(a) + f'(a)h)|}{h} = 0$$

- The function $df_a(h) = f(a) + f'(a)h$ is linear
- df_a approximates f at a and is called the differential
- The derivative is just the slope of the differential
- The key point: $|f(a+h) - df_a(h)| = o(h)$
- The speed at which h goes to zero is linear
- So the difference between f and its differential goes to zero **faster than linearly**
- Thus we can think of df_a as the linear part of $f(x)$ at a

From Derivatives to Taylor's Theorem

- We can restate the result of the last slide as follows

$$f(a+h) = f(a) + f'(a)h + o(h)$$

- Thus the differential relates the derivative to the best linear approximation to f at a point
- Taylor's theorem generalizes this relationship
- If f is k-times differentiable, then

$$f(a+h) = \sum_{j=0}^{k} f^{(j)}(a)h^j + o\left(h^{k+1}\right)$$

- Again, the most important part is the $o(h^{k+1})$ term
- Infinitely many polynomials are equal to f at a
- Only the Taylor polynomial has error of higher order

Multi-Dimensional Optimization

- Suppose we want to maximize or minimize $f(x)$
- If $x \in \mathbb{R}$, we can find the points $f'(x) = 0$
- Second derivative $f''(x)$ identifies maxs, mins, saddle points
- If $x \in \mathbb{R}^d$, we work with the **gradient**

$$\nabla f(x) = \left(\frac{\partial}{\partial x_1} f(x), \ldots, \frac{\partial}{\partial x_d} f(x) \right)$$

- Critical points $\nabla f(x) = 0$
- Need to check **Hessian** $H_f(x)$ to identify extrema

Constrained Optimization

- Sometimes we want to solve a problem of the form

$$\max_x f(x) \text{ such that } g(x) = 0$$

- The equation $g(x) = 0$ defines the subset we search over
- For example, if

$$g(x) = \sum_{i=1}^{d} x_i^2 - r^2$$

  then $g(x) = 0$ is the sphere of radius r
- Looking at $\nabla f(x) = 0$ no longer works
- Maxima of f on $g(x) = 0$ may not be maxima over all space
- If $f(x, y) = x^2 + y^2$, then every (x, y) is a max on any circle
- But no (x, y) is a local or global max of $f(x, y)$ overall

Lagrange Multipliers

- Lagrange multipliers allow us to solve this problem!
- To understand Lagrange multipliers, we need some geometry
- Suppose that on $g(x) = 0$, f has a global maximizer x_0
- Suppose furthermore that $f(x_0) = a$
- Then the curves $f(x) = a$ and $g(x) = 0$ are tangent
  - The curves must touch (or $f(x_0)$ could not equal a)
  - What if $f(x) = a$ passes through $g(x) = 0$?
  - Then $f(x) = a$ touches $g(x) = 0$ at ≥ 2 points
  - But we assumed x_0 was a global maximizer...
  - So $f(x) = a$ must just touch $g(x) = 0$
  - I.e. the two curves are tangent

Lagrange Multipliers

- The gradient connects this geometry to calculus
- The gradient points in the direction of steepest ascent
- I.e. f increases fastest from any point in that direction
- And $-\nabla f(x)$ is the direction of steepest descent
- What if we travel perpendicular to $\nabla f(x)$?
- Intuitively, $f(x)$ should be constant in this direction
- So $\nabla f(x)$ is perpendicular to $f(x) = a$ at x_0
- And $\nabla g(x)$ is perpendicular to $g(x) = 0$ at x_0
- And $g(x) = 0$ is parallel to $f(x) = a$ at x_0...
- So $\nabla g(x_0)$ and $\nabla f(x_0)$ are parallel!
- This allows us to find x_0

Lagrange Multipliers

- Let $\mathcal{L}(x, \lambda) = f(x) - \lambda g(x)$
- This is called the Lagrangian
- The λ variable is called the Lagrange multiplier
- The solution to the constrained maximization

$$\max_x f(x) \text{ such that } g(x) = 0$$

satisfies

$$\frac{\partial}{\partial x_i}\mathcal{L}(x, \lambda) = 0 \text{ for all } i$$

$$\frac{\partial}{\partial \lambda}\mathcal{L}(x, \lambda) = 0$$

Lagrange Multipliers

$$\mathcal{L}(x, \lambda) = f(x) - \lambda g(x)$$

- The condition $\frac{\partial}{\partial x_i}\mathcal{L}(x, \lambda) = 0$ ensures

$$\nabla f(x) = \lambda \nabla g(x)$$

- This just states that the gradients are parallel
- Also explains the need for the λ multipliers
- The condition $\frac{\partial}{\partial \lambda}\mathcal{L}(x, \lambda) = 0$ ensures $\nabla g(x) = 0$
- This just states that we satisfy the original constraint
- Thus solving the Lagrangian ensures that we are within the constraint and that the level sets of g and f are tangent at the solution point
- This is what characterized the maximum over our constaint

Basic Probability

- Now we will cover some basic probability
- We will be interested in two kinds of random objects:
  - Random events (i.e. whether it will rain tomorrow)
  - Random variables (i.e. a stock's price in a week)
- For events A and B, we define
  - $A \cap B$ to mean A and B happen (intersection)
  - $A \cup B$ to mean A or B (or both) happen (union)
  - $\neg A$ to mean A does not happen (negation)
  - $A \subset B$ to mean if A happens, B happens (inclusion)
- We define two special events: Ω and $\emptyset$
  - Ω is the union of all events: "something happens"
  - $\emptyset$ is the intersection of all events: "nothing happens"
- For an event A, $\mathbb{P}(A)$ denotes the probability of A
  - We define $\mathbb{P}(A)$ to be between 0 and 1
- A and B are disjoint if $A \cap B = \emptyset$ ("A and B can't both happen")

The Probability Axioms

- Probability is defined by the following rules:
  - $0 \leq \mathbb{P}(A) \leq 1$ for all events A
  - $\mathbb{P}(\Omega) = 1$ - "something always happens"
  - $\mathbb{P}(\neg A) = 1 - \mathbb{P}(A)$ - "A either happens or it doesn't"
  - If A and B are disjoint, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$
  - If $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$
- These rules have many consequences
- For example, $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$
  - Let $A \setminus B$ be the event A happens but B does not
  - Observe that

$$\mathbb{P}(A \cup B) = \mathbb{P}((A \setminus B) \cup B) = \mathbb{P}(A \setminus B) + \mathbb{P}(B) \leq \mathbb{P}(A) + \mathbb{P}(B)$$

Conditional Probability and Independence

- Conditional probability is an extremely important concept
- Let A and B be two events
- We define the probability of B given A as

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$$

- I.e. the proportion of cases in which A happens where B also happens
- Suppose B is an event that comes after A in time
- Before A, we expect B to happen with probability $\mathbb{P}(B)$
- But before B occurs, A either happens or it doesn't
- This new information might be relevant to B
- We now expect B to happen w/ probability $\mathbb{P}(B \mid A)$
- This kind of scenario comes up often in statistics

Conditional Probability and Independence

- We say that two events A and B are independent if

$$\mathbb{P}(A \mid B) = \mathbb{P}(A)$$

- This is equivalent to $\mathbb{P}(B \mid A) = \mathbb{P}(B)$
- "knowing that A happened gives no information about B"
- And vice versa
- Rearranging the definition of conditional probability:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B \mid A) = \mathbb{P}(B)\mathbb{P}(A \mid B)$$

- So if A and B are independent, we have

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

- This is in fact equivalent to independence

Bayes Rule

- If we combine the facts that

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \text{ and } \mathbb{P}(A \cap B) = \mathbb{P}(B \mid A)\mathbb{P}(A)$$

  we get

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

- This is known as Bayes rule
- It often comes in an alternate form
  - Rewrite the denominator as

$$\mathbb{P}(B) = \mathbb{P}(B \cap A) + \mathbb{P}(B \cap \neg A) = \mathbb{P}(B \mid A)\mathbb{P}(A) + \mathbb{P}(B \mid \neg A)\mathbb{P}(\neg A)$$

  - The verbose form of Bayes rule is then:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B \mid A)\mathbb{P}(A) + \mathbb{P}(B \mid \neg A)\mathbb{P}(\neg A)}$$

Bayes Rule

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B \mid A)\mathbb{P}(A) + \mathbb{P}(B \mid \neg A)\mathbb{P}(\neg A)}$$

- The verbose form is a very useful formula
- We can express one conditional in terms of the opposite
- Often one form is much easier than the other
- Causation only goes one direction
- What is the probability of having a disease given a positive test?
- We can compute this if we know:
  - The probability of a positive test given disease
  - The probability of a positive test given no disease
  - The probability of disease overall
- These are often much more readily known!

Random Variables

- In addition to events, we can consider random variables
- These represent random about which we are uncertain
  - Future values we cannot yet know
  - Measurements that cannot be made exactly
  - Unobservable quantities defined by theory
- We will define random variables to be random numbers
- Possible to define more complex random objects
- If X and Y are random variables, then
  - $\{X = a\}$
  - $\{a \leq X \leq b\}$
  - $\{f(X) \leq g(Y)\}$

  are (random) events
- If X is a real number, we say it is continuous
- If X is an integer, we say it is discrete

Random Variables and Distributions

- If X is a random variable, we don't know its value
- But we may know something about the process that generates X
- For example, we don't know if it will rain tomorrow
  - But we know about previous weather conditions
  - Have a sense of when rain is likely and when it isn't
- May know certain values of X are more likely than others
- This defines a **probability distribution** for X
- Distributions tell us about the probability that a random variables takes certain values

Discrete Distributions

- If X is discrete, then it's value will be an integer
- For each integer i, we can ask for $\mathbb{P}(X = i)$
- The function $p(i) = \mathbb{P}(X = i)$ is the probability mass function for X
- This function defines the distribution of X
  - I.e. it defines all possible probabilities for X
- For example:

$$\mathbb{P}(0 \leq X \leq 10) = \sum_{i=0}^{10} \mathbb{P}(X = i) = \sum_{i=0}^{10} p(i)$$

- Fundamental property of probability mass functions:

$$p(i) \geq 0 \text{ and } \sum_{i=-\infty}^{\infty} p(i) = 1$$

Continuous Distributions

- If X is continuous, it is some real number
- No longer makes sense to compute $\mathbb{P}(X = x)$
  - This probability must be zero!
- What would the continuous analog of a pmf be?

$$p(x) \geq 0 \text{ and } \int_{-\infty}^{\infty} p(x)dx = 1$$

- Such a function is a probability density function
- A density function also completely characterizes a distribution
- For example:

$$\mathbb{P}(a \leq X \leq b) = \int_{a}^{b} p(x)dx$$