IST557 Lecture Notes

Andrew Sugarman

September 3, 2024

Contents

1 08-27-2024 - Class Introduction

1.1 Syllabus

- all homework based unless people cheat
  - if you work together on homeworks you must report it and midterm is absolutely to be done alone
- only hw + take home
- TA office hours 2-4pm mon and 9-11 Fri

1.2 Background

- review cholesky factor
  - spend a fair bit of time over the next few days looking over linear algebra review and reference
  - dont cheat
  - For grading = understand the question
    - more credit if you recognize the answer is wrong
    - **no late assignments**
    - murphy and elements of stat learning = best books
      - Review of math and probability = high-yield

1.3 Lecture 1

Machine Learning Overview and context

- inference = you care about what the model has learned
  - supervised learning = predict Y from X given training examples where both were known ie housing price pred problem
  - unsupervised = predict y from x given examples where only x is known ie identify 5 groups from a dataset
  - semi-supervised learning = y is only known for part of the training data
  - Regression = continuous Y, classification = discrete Y
  - Feature selection
    - given Ys and Xs, figure out which covariates are the most important
  - Dimensionality reduction ex PCA

Model representations

- Data do not fall on a line
- Linear Regression
  - Probabilistic representation: Beta = vector containing m and b Beta = maximum likelihood of Y given beta*x and variance
    - maximize prob of the data under this model
  - Loss representation
    - find values of m and B that minimize the sum of the squared res

Black box model

- subjective label as to whether you can understand how it is working (nnet, random forest, decision tree)

Model evaluation

- are models any good? How do we define what good is?

P»N problem = more parameters than data points

- use penalizated or bayesian regression to help solve this

Stats vs ml

- stats generally focuses on inference

Types of Data

- discrete data with more than 2 categorical levels (one of K categories)
  - one hot encoding - 000, 100, 010, 001
  - dumy encoding = new variable z that is categorical w k-1 dims
- ordinal data - categorical with an order
- interval data - protect identity
- time to event data - how long to develop a condition of interest
  - special + complexities with specialized models
- Functional data (inf dim) measure continuous functions such as ekgs
- compositional data (sum constraints)

2 08-29-2024 - Math and Probability Review

2.1 Linear Algebra

$A_{i,j}$ means element from ith row and jth column

- can only add matrices of same dimensions
- can multipy two matrices that do not have same dim
- For A*B, the num columns (m) in A must be same as num rows (n) in B
  - The inner dimensions cancel out
- Matmul is associative, distributive, and not commutative
- every linear transformation can be represented by a matrix

ONLY square matrices could be invertible, and not all matrices that are square have a unique inverse

- $AA^-1 = A^{-1}A = I$
- pseudoinverse = $\mathrm{A}^{\mathrm{dagger}}$: defined by $A^t * A = I$
  - there is not a unique solution to the inverse
    - pseudoinverse in python gives arb value for Adagger
      - this means that sometimes there are problems with random answers if code is using pseudoinv
      - R does not give you an answer if the inverse is not defined
      - assumptions can help ie most betas are small

Identity matrix

- zero except for diagonal of ones

Diagonal matrix

- usually only well-defined for square matrices Diag(X) –> shorthand notation for either extracting or creating a diagonal matrix

Special matrices

- symmetric = equal to the transpose (such as a covariance mat)
- orthogonal = things that rotate or translate vectors but do not scale them
  - the inv of an ortho matrix is its transpose

Linear dependence

- 3 matrices on the same plane are linearly dependent and the matrix with these three vectors as rows would have rank 2
- span(S) is the set of all linear combinations of the elements of S

Rank

- AR^mxn
- Rank(A) is the max num of linearly ind columns or rows

Eigenvectors and Eigenvalues

- eigenvectors are usually normalized to unit length
- if A is symmetric then all eigenvalues r real
- trace of a matrix = sum of the eigenvals
- det(A) = product of eigenvals
- If X = VDV^T then: $X^-1 = V * D^-1 * V^T$
  - since D is diagonal its inverse is given by just taking the inverse of each of its diag elements
  - this is beneficial because matrix inversion is computationally expensive

A symmetric Positive definite matrices have all eigenvalues strictly greater than 0 A symmetric matrix is called pos semi definite if all eigvals are greater than 0 (but can include 0) = covariance matrices Matrix Square roots

- the square root of a square matrix X is defined as mat V such that X = VV^T
  - eigen decomp provides means of finding such a mat V for sq mat X
    - $X^{1/2} = VD^{1/2}$

IF X is spd (symmetric pos def) Cholesky decomp is faster than eigen decomp

- SPD matrix sigma has chol decomp: Sigma = LL^T where L is a lower triangular matrix
  - TLDR Cholesky decomp for sigma = LL^T if sigma is a symmetric positive definite matrix
    - upper cholesky = U^T * U

Vector norms Think of a norm as the length of a vector

1. Euclidean norm = $||x_2||$ = $\sqrt{sumx^2}\backslash$
2. L1 norm = city block norm (Lasso)
3. p-norm = pthroot(sum of absval x^p)

Recall derivative = slope of tangent line, inst rate of change at a point The gradient = multivariate derivative

- for fn F that takes in a vector and outputs a scalar (such as a probability)
  - gradient is defined as a vector
  - nobla*f =
  - gradient points in the direction of steepest ascent from x and -nobla(f(x)) gives direction of steepest descent
  - this is used frequently in gradient descent

Jaccobian is a matrix of first order partial derivatives (when the output is a vector) ie a generalization of the multivariate gradient Hessian is a matrix of second order partial derivatives

- think of this as the curvature of a function
- comes in handy for newtons method
- serves also as the basis for the Laplace Approximation to a probability density

Review of Optimization

3 09-03-2024 - Math and probability review

3.1 The Eigendecomposition is ordered

- first eigenvector has the greatest value

3.2 Argmax

- For a function

3.3 Joint probability (memorize this)

- factor into a conditional and a marginal
- $P(A|B) = P(A|B)(P(B)) = P(B|A)(P(A))$

3.4 Conditional Probability

$p(B|A = 7) = P(A = 7, B)/P(A)$

3.5 Expectations, the Mean, and variance

- expectation = weighted avg

Variance is the spread about the mean - it must be positive

3.6 Mean and variance of finite samples

$mean_(x)=1/N*\sum x_i$