Math 10B with Professor Stankova
Worksheet, Discussion #27; Tuesday, 5/1/2018
GSI name: Roy Zhao

1 Linear Regression and Correlation

1.1 Concepts

1. Often when given data points, we want to find the line of best fit through them. To them, we want to approximate them with a line $y = ax + b$. We represent this as a solution where we want to solve for a, b. In matrix vector form and data points (x_i, y_i), this is represented as
$$A\vec{x} = \vec{b} \to \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$
Often, we cannot find a perfect fit (if not all the points lie on the same line). So we want to find the error. One way to find the error is to take the least square error or $E = \sum (y_i - (ax_i + b))^2$, the sum of the squares of the error. The choice of a, b that minimizes this is
$$\begin{pmatrix} a \\ b \end{pmatrix} = (A^T A)^{-1} A^T \vec{b}.$$
Written out, we have
$$a = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x - \bar{x})^2}, b = \bar{y} - a\bar{x},$$
where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ is the average of the x values and $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ is the average of the y values.

   The **correlation coefficient** of a set of points $\{(x_i, y_i)\}$ is given by
$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}.$$
Another way to represent that the correlation coefficient is the cosine of the angle between the two vectors $\vec{x} = (x_i - \bar{x})$ and $\vec{y} = (y_i - \bar{y})$. So, we can write
$$r = \frac{\vec{x} \circ \vec{y}}{|\vec{x}||\vec{y}|}.$$
It is always between -1 and 1 by Cauchy-Schwarz.

1.2 Example

2. Find the line of best fit and correlation coefficient of the following data $\{(-2,-2),(2,3),(3,1)\}$.

Solution: We can use the formulas. First we calculate $\bar{x} = \frac{-2+2+3}{3} = 1$ and $\bar{y} = \frac{-2+3+1}{3} = 0$. This gives us that

$$a = \frac{(-2-1)(-2-0)+(2-1)(3-0)+(3-1)(1-0)}{(-2-1)^2+(2-1)^2+(3-1)^2} = \frac{11}{19}$$

and $b = \bar{y} - a\bar{x} = 0 - \frac{11}{19} = \frac{-11}{19}$. Therefore, the line of best fit is $y = \frac{11}{19}x - \frac{11}{19}$. The correlation coefficient is

$$r = \frac{(-2-1)(-2-0)+(2-1)(3-0)+(3-1)(1-0)}{\sqrt{(-2-1)^2+(2-1)^2+(3-1)^2}\sqrt{(-2-0)^2+(3-0)^2+(1-0)^2}} = \frac{11}{\sqrt{19}\sqrt{14}}.$$

$r \approx 0.6745$.

1.3 Problems

3. True **FALSE** The line of best fit always exists.

Solution: If there is only one point or all the points have the same x value, then the line of best fit will not exist.

4. **TRUE** False The matrix A^TA will always be square.

Solution: If A is $m \times n$, then A^T is $n \times m$ so A^TA is $(n \times m)(m \times n) = n \times n$ so it will be square.

5. **TRUE** False The correlation is always between -1 and 1.

6. True **FALSE** If the correlation between two sets of data is -1, then y is proportional to x^{-1}.

7. **TRUE** False If we shift the data (by for instance adding 5 to all of the y values), then the correlation does not change.

8. Consider the set of points $\{(-2,-1),(1,1),(3,2)\}$. Calculate the line of best fit and the correlation (no need to simplify calculations)

Solution: We first write it in matrix form as

$$\begin{pmatrix} -2 & 1 \\ 1 & 1 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \\ 2 \end{pmatrix}.$$

Then we calculate $A^TA = \begin{pmatrix} 14 & 2 \\ 2 & 3 \end{pmatrix}$ and $A^T\vec{b} = \begin{pmatrix} 9 \\ 2 \end{pmatrix}$. Then

$$\begin{pmatrix} a \\ b \end{pmatrix} = (A^TA)^{-1}A^T\vec{b} = \frac{1}{42-4}\begin{pmatrix} 3 & -2 \\ -2 & 14 \end{pmatrix}\begin{pmatrix} 9 \\ 2 \end{pmatrix} = \begin{pmatrix} 23/38 \\ 10/38 \end{pmatrix}.$$

So the line of best fit is $y = 23/38x + 5/19$.

The correlation is

$$r = \frac{(-2-2/3)(-1-2/3)+(1-2/3)(1-2/3)+(3-2/3)(2-2/3)}{\sqrt{(-2-2/3)^2+(1-2/3)^2+(3-2/3)^2}\sqrt{(-1-2/3)^2+(1-2/3)^2+(2-2/3)^2}} = \frac{69}{\sqrt{114}\sqrt{42}}.$$

$r \approx 0.9972.$

9. Consider the set of points $\{(-2,-1),(1,1),(3,2)\}$. Calculate the square error if we estimate it using the line $y = x$. Then calculate the square error if we use the line $y = 0$. Which is a better approximation?

Solution: Using the fit $y = x$ gives the error $(-1-(-2))^2+(1-1)^2+(2-3)^2 = 1+0+1 = 2$. Using the approximation $y = 0$ gives the error $(-1-0)^2+(1-0)^2+(2-0)^2 = 1+1+4 = 6$. Thus $y = x$ is the better fit.

10. Find the line of best fit and the error of the fit of the points $\{(-1,2),(0,-1),(1,1),(3,2)\}$ and use it to estimate the value at 2. Calculate the correlation of the data.

Solution: We first write it in matrix form as

$$\begin{pmatrix} -1 & 1 \\ 0 & 1 \\ 1 & 1 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \\ 1 \\ 2 \end{pmatrix}.$$

Then we calculate $A^TA = \begin{pmatrix} 11 & 3 \\ 3 & 4 \end{pmatrix}$ and $A^T\vec{b} = \begin{pmatrix} 5 \\ 4 \end{pmatrix}$. Then

$$\begin{pmatrix} a \\ b \end{pmatrix} = (A^TA)^{-1}A^T\vec{b} = \frac{1}{44-9}\begin{pmatrix} 4 & -3 \\ -3 & 11 \end{pmatrix}\begin{pmatrix} 5 \\ 4 \end{pmatrix} = \begin{pmatrix} 8/35 \\ 29/35 \end{pmatrix}.$$

So the line of best fit is $y = 8/35x + 29/35$. Thus, $y(2) = 8/35(2) + 29/35 = (16+29)/35 = 45/35 = 9/7$.

The correlation is

$$r = \frac{(-1-3/4)(2-1) + (0-3/4)(-1-1) + (1-3/4)(1-1) + (3-3/4)(2-1)}{\sqrt{(-1-3/4)^2 + (0-3/4)^2 + (1-3/4)^2 + (3-3/4)^2}\sqrt{(2-1)^2 + (-1-1)^2 + (1-1)^2 + (2-1}}$$

$$= \frac{2}{\sqrt{35/4}\sqrt{6}} = \frac{4}{\sqrt{35}\sqrt{6}} \approx 0.2760.$$

2 Review Topics

2.1 Counting and Probability

- Permutations and Combinations
  - Binomial coefficients
  - Poker-type problems
- Principle of Inclusion-Exclusion (PIE)
  - Complementary Counting
- Pigeonhole principle
- 12-Fold way
  - (In)distinguishable balls and (in)distinguishable boxes
  - Sterling Numbers of the Second kind
  - Partition Numbers
- Probability, Expected Value, Variance, Covariance
  - Random variable picture
  - Probability Mass Functions (PMF)
  - Formulas for expected values
  - Conditional probability

  - Independence

- Bayes' Theorem
- Distributions
  - Uniform Distribution
  - Bernoulli Trials
  - Binomial distribution
  - Hyper-geometric distribution
  - Geometric distribution
  - Poisson distribution
  - Normal distribution
  - χ^2 distribution
  - Expected value and variance of each
- Hypothesis Testing
  - Central Limit Theorem Testing
    - Z-Scores
  - χ^2 Testing
  - Independence Testing
  - Null/Alternative Hypotheses
  - Type 1/type 2 errors, significance level, power
- Estimators and Confidence Intervals
  - Estimators for the mean and standard deviation
  - 95% confidence intervals

2.2 Differential Equations

- Recurrence Relations
  - Going both forward and backward
  - Writing one in matrix form
  - Finding formulas for first order linear equations
- Identifying the adjectives (linear, homogeneous, etc.)
- Integrating Factors
- Separable Equations

  - Logistic Growth
  - Exponential Growth
  - Partial Fractions

- Second order differential equations
  - Going forward and backward
  - Writing one as a system of linear first order equations

- IVPs/BVPs

- Slope fields
  - Euler's Method

- Linear systems of differential equations

2.3 Matrices

- Multiplying matrices, vectors

- Cauchy-Schwarz Inequality

- Determinants
  - Number of solutions and how it depends on the determinant

- Gaussian Elimination
  - Consistent vs Inconsistent systems
  - Finding Inverses
  - Solving matrix-vector equations

- Eigenvalues/eigenvectors

- Linear Regression
  - Least Squares Error
  - Finding line of best fit

- Correlation

2.4 Miscellaneous

- Euler's Formula
- Induction
- Sorting Algorithms
  - Bubble Sort
  - Quick Sort
- Stable-matching algorithm

True/False

1. **TRUE** False Changing the initial conditions for a linear homogeneous recurrence relation does not affect the bases of the exponential functions that appear the direct formula for the relation.
2. **TRUE** False The difference operator $\triangle$ takes a sequence and makes a new sequence out of it.
3. **TRUE** False The DE $y'(t) = 0.04\,y(t) - 0.09$ can be solved in at least 2 different ways.

Solution: We can use separable equations and integrating factors.

4. True **FALSE** Checking that a function $y(t)$ is a solution to a DE may not be possible since we may not know how to solve the DE.
5. True **FALSE** The equation $e^x y' = y$ is linear, but $y' + x^2 e^x y = e^x$ is not linear.

Solution: Both are linear.

6. **TRUE** False There are IVP's in which the function $f(t, y)$ is continuous everywhere, but the solutions to the IVP cannot extend beyond a certain interval $[0, T)$.
7. **TRUE** False An ODE is both linear and separable exactly when it is of the form $y' = (y + c)\,g(t)$ for some function $g(t)$ and some constant c .
8. **TRUE** False Solutions to a separable ODE can "go missing" when both sides of the ODE are divided by a function of y.

9. True **FALSE** All linear ODE's have the property that linear combinations of their solutions are also solutions to them.

Solution: All linear **homogeneous** ones have this property.

10. **TRUE** False All I.V.P.'s for second order, linear, homogeneous ODE's with constant coefficients are solvable and have a unique solution.

11. **TRUE** False Some B.V.P. for second order, linear, homogeneous ODE's with constant coefficients may have no solutions, a unique solution, or infinitely many solutions, but never any other number of solutions (e.g., exactly 2 solutions).

12. True **FALSE** The DE $y' = 3y^2$ will have a slope field with same slopes lined up in vertical lines because the equation is autonomous.

Solution: They will have the same slopes lined up in horizontal lines because it does not depend on t.

13. **TRUE** False If y_1 and y_2 are solutions to $y'' - 6y' + 5y = 4x$, then $3y_1 - 2y_2$ is also a solution to the DE.

14. **TRUE** False A vector can be represented algebraically as a $1 \times n$ or an $n \times 1$ matrix.

15. True **FALSE** The dot product of vectors always yields a non-negative result, but it is the norm of a vector that gives its length.

Solution: The dot product may be negative.

16. True **FALSE** Two vectors (of same dimensions) are perpendicular if and only if their dot product is 1.

Solution: They are perpendicular if their dot product is 0.

17. **TRUE** False For any two non-zero vectors $\vec{v}_1$ and $\vec{v}_2$ in the plane (of dimensions 2×1), we can find the angle α between them by the formula: $\alpha = \arccos \frac{\vec{v}_1 \circ \vec{v}_2}{|\vec{v}_1| \cdot |\vec{v}_2|}$

18. True **FALSE** An invertible matrix A could be of any size, even non-square, as long as its product with its inverse matrix equals the identity matrix.

Solution: An invertible matrix must be square.

19. **TRUE** False For any matrix $A_{m\times n}$ there is another matrix $B_{m\times n}$ such that $A + B = 0m \times n$ and this matrix B is unique.

20. True **FALSE** Diagonal matrices are the only matrices that equal their own transposes.

21. **TRUE** False There are non-square matrices A and B for which it is possible to multiply them in either order but then AB cannot equal BA.

22. **TRUE** False The determinant of a 2×2 matrix A determines whether the system $A\vec{x} = \vec{b}$ will have a unique solution or not, but it cannot distinguish by itself between systems with no solution and with infinitely many solutions.

23. True **FALSE** The system $D\vec{x} = \vec{b}$ where D is a diagonal matrix will have a unique solution exactly when D has a zero entry along the diagonal.

24. True **FALSE** To find the inverse A^{-1} of a square matrix A by Gaussian elimination, we reduce the "double matrix" $(A|I_n)$ by elementary row operations to $(U|A^{-1})$ for some upper-triangular matrix U.

25. **TRUE** False The determinant of an upper-triangular matrix U is equal to the determinant of a diagonal matrix D with same entries as U along the diagonal.

26. True **FALSE** As soon as we see a row like $(000\dots0|0)$ during Gaussian elimination, we know that the system will have infinitely many solutions.

27. True **FALSE** An eigenvector can be the zero-vector but an eigenvalue cannot be 0.

28. **TRUE** False When applying the algorithm to search for eigenvectors of a matrix, we must first find the eigenvalues, even if the problem is not asking for them.

29. True **FALSE** If an eigenvector $\vec{v}$ for a matrix A corresponds to eigenvalue $\lambda = 2018$, then $A^{2019}(\vec{v}) = 2019^{2018}$

30. **TRUE** False The reason that we set det $B = 0$ where $B = A - \lambda I$ is to ensure that the system of equations $B\vec{x} = \vec{0}$ has more than just the trivial solution.

31. True **FALSE** No matter what type of problem we are asked to solve, we can always skip finding the eigevectors and get by with just the finding the eigenvalues.

32. True **FALSE** A^TA cannot be a symmetric matrix if A is not square.

33. True **FALSE** The sum of the residuals of data points from a line is not a good estimate of the fitness of the line, since this sum could be large, yet the data points could be very close to the line.

34. True **FALSE** The least-square best-fitting line for any number of data points always exists and is unique essentially because there is a (unique) shortest distance from a point to a plane in any dimensions.

35. **TRUE** False If we use more data points to find the best-fiting line, we may increase the overall error S yet still be able to make better predictions about the data.

36. **TRUE** False The correlation is a number that is always between -1 and 1.