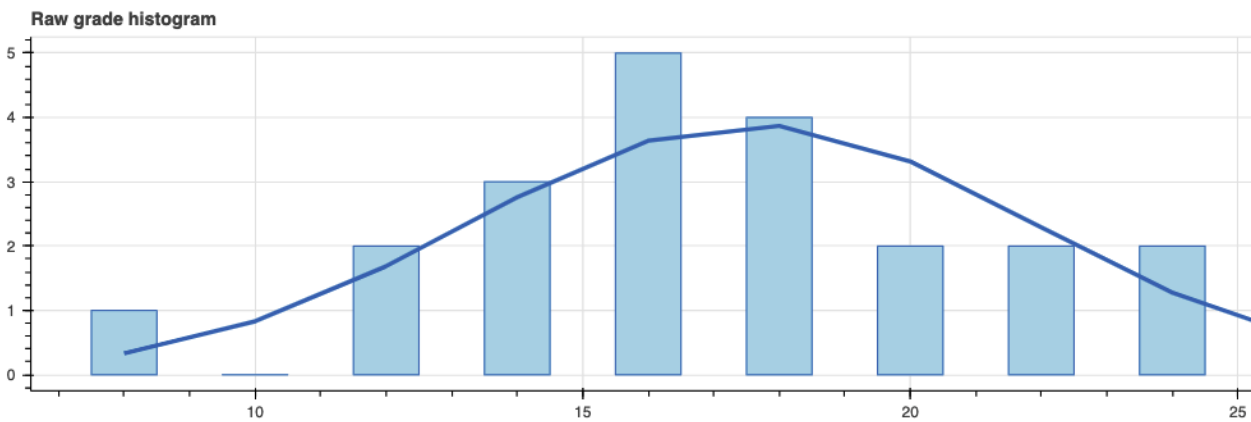


Class 7 overview: 7 March

Announcements & Carry-forwards

- 1. Performance on the midterm.



Class survey: Should we organize into small (3 person) groups for projects?

First hour


- Questions from the midterm,
- HW review (might not take a full hour)
- Statistics Unit 2 Linear regression - variance analysis


Second hour

- Gram-Schmidt example: "DC cherry blossom data".

Community Connection: Cherry Blossom Bloom Dates in Washington, D.C. | US EPA

This feature tracks the annual peak bloom date of Washington, D.C.'s famous cherry trees.

 https://www.epa.gov/climate-indicators/cherry-blossoms?utm_source=newsletter&utm_medium=email&utm_campaign=newsletter_axiosam&stream=top



U.S. ENVIRONMENTAL PROTECTION AGENCY

Third hour

- Review Gram Schmidt programming
- (time allowing) Statistics Unit 3 - "inverse probability" & Posterior inference.

Fourth hour

- Introduction to NLP sentence embedding data set.

Assignments

Linear regression & design matrices:

- Derive the "normal equations" for a *univariate* (just one explanatory variable) linear regression, to solve for the beta parameters by minimizing mean squared error, $\sum_i^n (y_i - \hat{y}_i)^2$:

$$y = \beta_0 + \beta_1 x + \epsilon$$

- **“One-hot” encoding.** One common technique to make predictions from unordered, categorical variables (e.g. gender, country, genre) is to create a separate column in the design matrix for each category, and code an indicator variable, for whether that category (column) applies to that sample (row)
- Draw out an example for the “one-hot” expansion for a two category regression, including an ‘intercept’ column.
- Why might problems with the math arise in solving this linear regression? Perhaps run an example to test your intuition.

Programming: Sentence Embedding exploration.

This will be the first task in your class project over the next 4 weeks.

The IMDB database can be used to predict movie sentiment - a binary “positive” or “negative” indication. One converts the review text to embedding vectors using the sentence transformers, (See the notebook *sentence_transformers.ipynb*) then uses them as features to predict the sentiment category.

The challenge is that such a regression is unwieldy and inefficient. One reason is that some of the columns created from the features of the embeddings may be duplicative or irrelevant. The goal is to come up with a more predictive set of columns


- First step is to load the small dataset,

<https://archive.ics.uci.edu/dataset/331/sentiment+labelled+sentences>

- convert it into sentence embeddings and run a regression on it, to set a baseline for performance. Please use the python statsmodels regression api to generate extensive regression diagnostics.
- The approach, roughly, is to select a set of columns that improve the prediction. Think how one can use a similarity test based on inner products for this, and see if you can (by trial and error) improve the results.
- We will need to come up with a criterion to measure performance. Since performance will have a probability distribution, we will need to design a statistical test for the criterion.

Files

- Mid-term solutions (revised). see folder “class_five”
- Statistics_Unit2
- CherryBlossomPeak.ipynb
- cherry_blossoms_fig-1.csv
- sentence_transformers.ipynb
- IMDB movie review database see

UCI Machine Learning Repository
Discover datasets around the world!
 <https://archive.ics.uci.edu/dataset/331/sentiment+labelled+sentences>

Suggested web slides

This lecture from U Washington covers much of the same class material: (It’s under “Files).

07_BasisFunctions.pdf

Additional Graphics

4 Basic Principles of Engaging Small Group Instruction

- P** **Positive Interdependence:**
"Does one doing well help others?"
"Does task completion depend on everyone doing his/her part?"
- I** **Individual Accountability:**
"Must everyone perform in front of someone?"
- E** **Equal Participation:**
"Is participation approximately equal? Time? Turns?"
- S** **Simultaneous Interaction:**
"What percent are performing at any one moment?"

How is ML different than conventional statistics?

