Class 8 overview: 14 March

Announcements & Carry-forwards

1. Plan for the 3 Project assignments
2. Next two classes cover regularized regression, bootstrap, significance testing, ROC curves

First hour

- HW review
- Probability unit 5: Entropy and measures of association. (runs into second hour)

Second hour

- Linear Algebra unit 6: Eigenvalues, SVD and PCA

Third hour

- Feature reduction & PCA. embedding_vector_PCA notebook.
- (Possibly presentation of temporal anomaly detection by PCA)

Fourth hour

- Class programming assignments on NLP data
- Description of project assignments.

Assignments

Entropy

1. Use the definition of Entropy and mutual information:

$$MI(Q;P) = H(P) - H(P \mid Q)$$

to derive:

$$MI(Q;P) = -\sum_i \sum_j P(x_i, y_j) \log[\frac{P(x_i)P(y_j)}{P(x_i, y_i)}]$$

Note the minus sign. How would the formula change without the minus sign?

Linear Algebra unit 6

1. Show that two matrices A, B that commute, e.g. $AB = BA$, share the same eigenvectors.
2. If a square matrix has a non-zero nullspace show that it has non-zero eigenvalues with eigenvalues == 0.
4. As noted, “too many” columns in the regression design matrix can lead to problems and unneeded computation. Alternately adding new features to create new columns has diminishing value. In the worst case adding duplicate columns creates a singular Gram matrix.
    a. What about duplicating rows in the design matrix. How does that affect the regression results?
    b. What is the effect on the prediction if a copy of all the rows is appended to the design matrix?
    c. What is the effect on prediction if a selected set of rows is appended to the design matrix?

First project assignment - build on what you have from the previous assignment.

Programming: Sentence Embedding exploration.

This will be the first task in your class project over the next 3 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 on Google Drive *HW7/*sentence_tranformers.ipynb)* then use them as features to predict the sentiment category.

1. 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+sente
```

- convert it into sentence embeddings and run a regression on it, to set a baseline for performance. See the notebook *HW/embedding_vector_PCA.ipynb. Please use the python statsmodels regression api to generate extensive regression diagnostics.

2. 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. *Can one use association measures to improve the regression?*

- How could one tell if the model gives better predictions? We will need to come up with a criterion to measure performance. **not just R^2, BIC or log-likelihood.*

3. Propose a measure of *stability* of the coefficients. Maybe if you subset the columns, see how much does it change the prediction?
4. ***Visualize your findings.***
5. (Thinking ahead for next week:) Since performance will have a probability distribution, we will need to design a statistical test for the criterion. *eg how would you put an interval around them?*

Files

- embedding_vector_PCA.ipynb
- IMDB movie review database see the link above
- Probability unit 5
- Linear Algebra unit 6

Suggested web slides

Watch this video on the Base Rate Fallacy, an application of Bayes rule.

The medical test paradox (well "paradox")

A link to the full video about Bayesian thinking is at the bottom of the screen.Or, for reference: https://youtu.be/lG4VkPoG3koLong-to-short editing by

https://www.youtube.com/shorts/xIMlJUwB1m8