AI Enterprise Workflow Study Group

Course 3, Week 1

3/21/2020

Agenda

- Check in
- Discussion
- Next steps

Course & Study Group Schedule

AI Enterprise Workflow Study Group		
Session	**Topic**	**Date**
Overview Webinar	Webinar with instructor, Ray Lopez	15-Feb
Course 1 Week 1	Course intro	22-Feb
Course 1 Week 2	Data ingestion, cleaning, parsing, assembly	29-Feb
Course 2 Week 1	Exploratory data analysis & visualization	7-Mar
Course 2 Week 2	Estimation and NHT	14-Mar
Course 3 Week 1	Data transformation and feature engineering	21-Mar
Course 3 Week 2	Pattern recognition and data mining best practices	28-Mar
Course 4 Week 1	Model evaluation and performance metrics	4-Apr
Course 4 Week 2	Building machine learning and deep learning models	11-Apr
Course 5 Week 1	Deploying models	18-Apr
Course 5 Week 2	Deploying models using Spark	25-Apr
Course 6 Week 1	Feedback loops and monitoring	2-May
Course 6 Week 2	Hands on with OpenScale and Kubernetes	9-May
Course 6 Week 3	Captsone project week 1	16-May
Course 6 Week 4	Captsone project week 2	23-May

Course 3 Week 1 learning objectives

1. Discuss feature engineering and transformations in the context of the AI workflow
2. Employ the tools that help address class and class imbalance issues
3. Explain the ethical considerations regarding bias in data
4. Employ dimension reduction techniques for both EDA and transformations stages
5. Describe topic modeling techniques in natural language processing
6. Use topic modeling and visualization to explore text data

Feature Engineering and Transformation

sklearn Interfaces:

- Transformer: Convert data from one form to another
- Estimator: Build and fit models
- Predictor: Make predictions

Class Imbalance

<table>
<tr><td></td><td></td><td colspan="2">True condition</td><td></td><td colspan="2"></td></tr>
<tr><td></td><td>Total population</td><td>Condition positive</td><td>Condition negative</td><td>Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$</td><td colspan="2">Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$</td></tr>
<tr><td rowspan="2">Predicted condition</td><td>Predicted condition positive</td><td>True positive</td><td>False positive, Type I error</td><td>Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$</td><td colspan="2">False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$</td></tr>
<tr><td>Predicted condition negative</td><td>False negative, Type II error</td><td>True negative</td><td>False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$</td><td colspan="2">Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$</td></tr>
<tr><td></td><td></td><td>True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$</td><td>False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$</td><td>Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$</td><td rowspan="2">Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$</td><td rowspan="2">$F_1$ score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$</td></tr>
<tr><td></td><td></td><td>False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$</td><td>Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$</td><td>Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$</td></tr>
</table>

Class Imbalance

Choosing appropriate loss function:

- Accuracy vs precision/recall or F1

Adjusting training data by up/over sampling or down/under sampling

- Downsampling easiest but lose data
- Many flavors of techniques for upsampling: SMOTE, ADASYN, SMOTENC, etc.
- NOTE: Split train/test sets first!

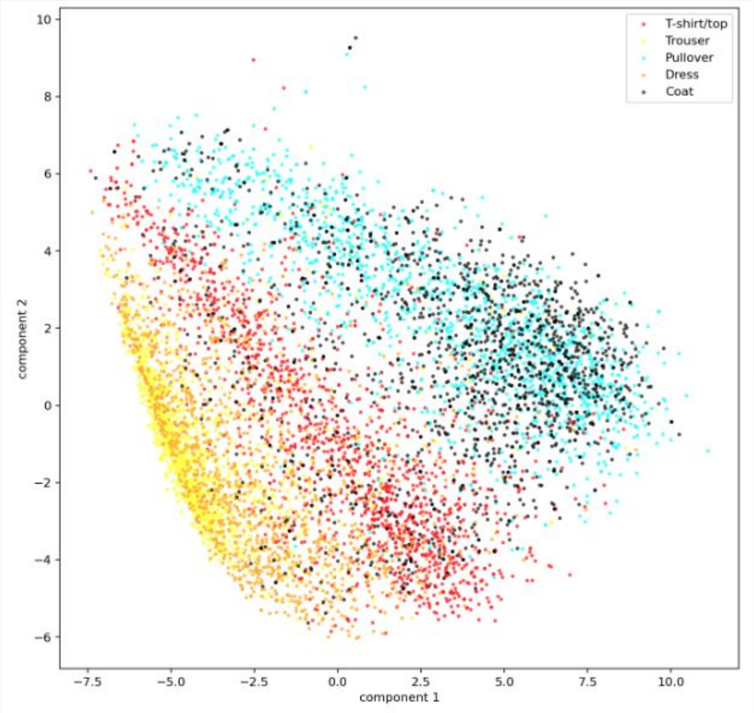
Neural nets very sensitive to imbalance. SVMs and tree methods more resilient.

Dimensionality Reduction

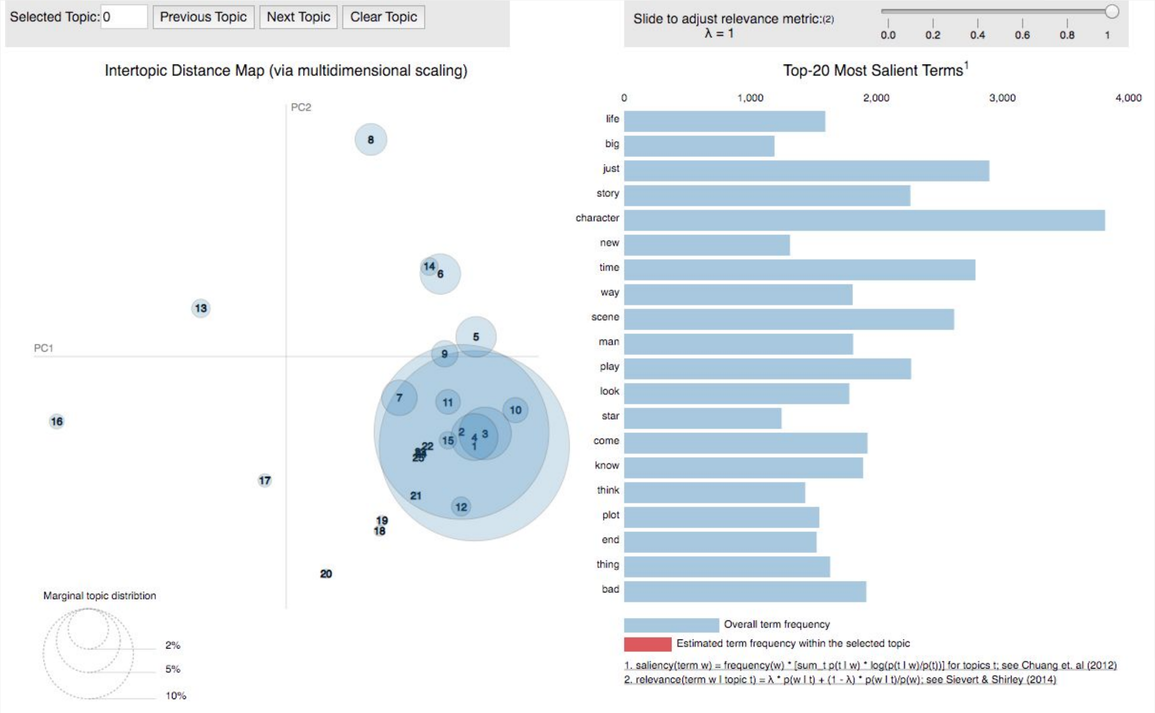
- Enable EDA visualization in high dimensional data
- Remove coliinearity
- Remove redundant features
- Deal w/ curse of dimensionality (statistical significance doesn't like sparsity, as the amount of data needed to support the result often grows exponentially with the dimensionality)
- Identify structure for supervised learning
  - Non-negative matrix factorization,
  - Autoencoders
  - Self-organizing maps

Applications: image analysis, text analysis, signal processing, astronomy, medicine, etc.

Fashion MNIST Example



Topic Modeling Case Study



Additional Discussion

What did you learn?

What stumbling blocks did you run into?

How do these lessons relate to your experience?

What did you learn/find interesting in this week's lesson?

What are you doing as homework?

What interesting resources have you found?

Other?

Next steps

Next week we move on to week 2 of Course 3, which continues the discussion of feature engineering with a focus on outlier detection and clustering.

Chime in on Slack if you run into any issues or want to share any observations

Prepare your questions, discussion points, etc. for next week's meetup