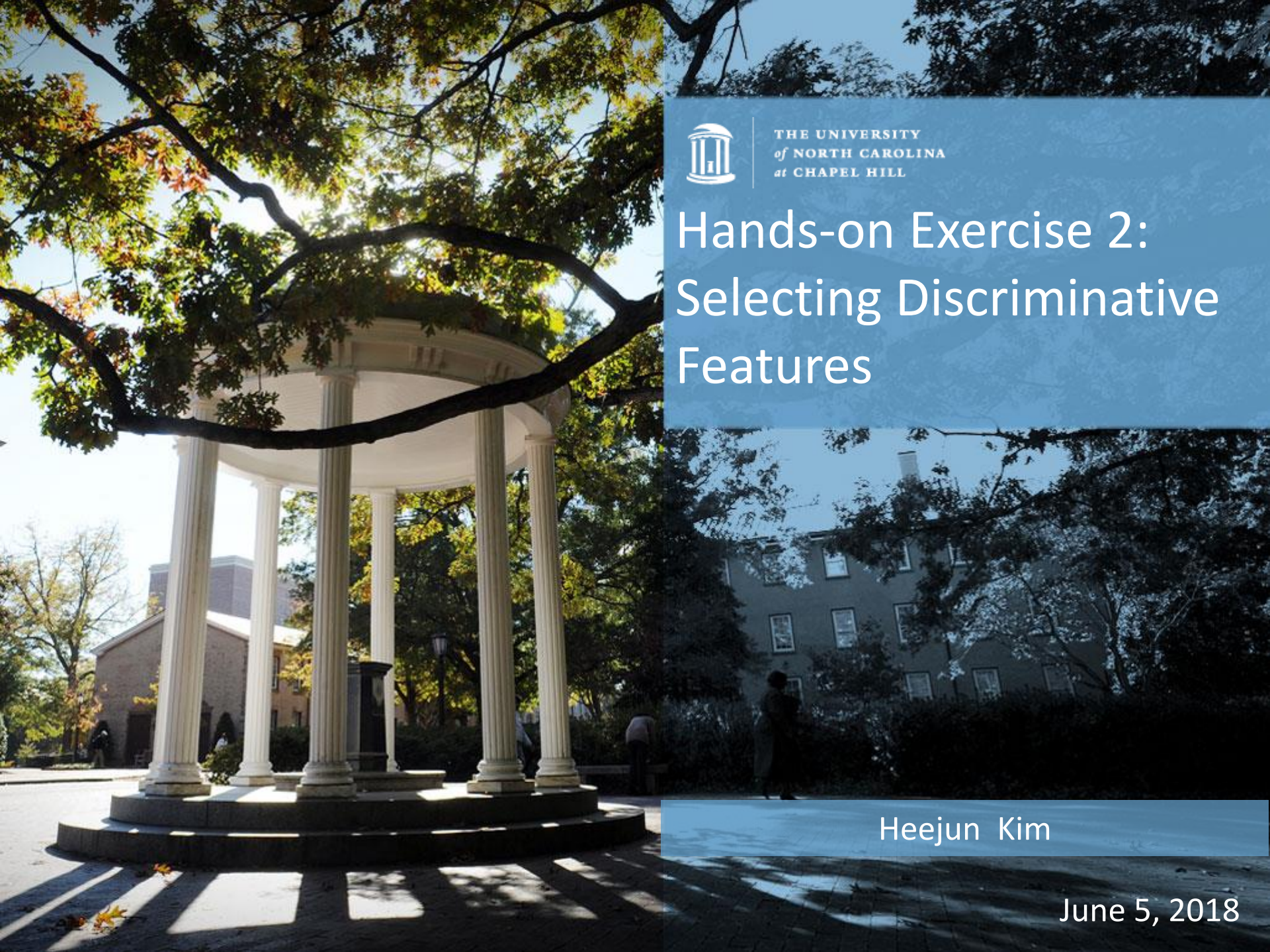

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Hands-on Exercise 2: Selecting Discriminative Features

Heejun Kim

June 5, 2018

Objective of the Exercise

- Develops a machine learning model to classify documents into breast neoplasms and skin neoplasms
- Producing and selecting features that are predictive of topics and improving the performance of the classifier is the objective of this exercise
- You are given a text file which includes 2 classes (training.csv and testing.csv)

Details of the exercise

- Due: June 12th (before class)
- What to do
  - Submit the screen shot of the best classifier from the LightSIDE (or other tool)
  - Save the trained model to test it with test data
  - Answer following questions in the class
    - Describe your strategy for improving the baseline classifier. What steps did you take in deciding what features to include/exclude
    - Using the predictions from your best classifier, find an interesting example of a false positive/negative mistake (explore results). Prepare this example in the slides. Explain why is this example interesting in the class. What kind of information would the classifier have to consider in order to avoid this particular type of mistake?

Details of the exercise

- How to improve
  - Explore different features (extract features)
  - Use metrics that measure the degree of co-occurrence between a feature and a target class value (extract features)
  - Play with the threshold (extract features)
  - Do error analysis (explore results)
  - Find external resources
- Rules
  - Only use Naïve-Bayes algorithm
  - Only use cross-validation as evaluation option

Any Questions?

Join the Piazza: http://piazza.com/unc/summer2018/hidav_text