Quantitative Text Analysis

Meeting 6

Petro Tolochko

Machine Learning

- Supervised
    - An outcome variable is defined
    - Focus is on prediction
- Unsupervised
    - No outcome variable has been defined
    - Focus is on patterns

Machine Learning

- **Supervised**
    - **An outcome variable is defined**
    - **Focus is on prediction**
- Unsupervised
    - No outcome variable has been defined
    - Focus is on patterns

Supervised

- Objective:
    - Classification of documents into pre existing categories

Supervised

- Create a labeled data set
- Classify documents with supervised learning algorithm
- Check performance

Labeled Dataset

- How:
  - Human coders annotate parts of the corpus (what we did together)
  - Found data (e.g., self-reported profession in users' profile)

- Considerations:
  - Sampling should be representative for the corpus (e.g., Random, Stratified sample e.g., across time and source)
  - Quality of human coding matters (Assess the intercoder reliability)
  - Number of documents

Labeled Dataset

- Number of documents
    - the higher the number of categories and the lower the reliability of the coders, the higher the number of documents (Barberá et al., 2021)
- increase the sizes of manually coded validation dataset as large as possible (e.g., more than 1% of all data to be examined), assuming acceptable reliability (equal to or higher than .7) (Song et al., 2021)

Splitting the Data

- Split labeled data in training data and test data (validation data)
- Training data
    - The subset that is used to learn the model parameters
- Test data
    - Another subset used to evaluate the model's predictive quality
    - Not used for learning!
- Validation data

Document Classification

- Classifier learns the mapping between features and the labels in the training set
- define a model Y=g(X)
- And apply a learning algorithm to establish which features in X (features extracted from the training documents) matter to recover Y (i.e, the labels of the training documents)
- We fit the model

Machine Learning 101

Machine Learning 101

- Model: $$\boldsymbol{Y} = f(\boldsymbol{X})$$

- Objective function (e.g.,): $$MSE = \frac{1}{n}\Sigma_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

- Optimisation: $$argmin_{\hat{Y}}\frac{1}{n}\Sigma_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Machine Learning 101

- Model:

$$Y = f(X)$$

- Objective function (e.g.,):

$$MSE = \frac{1}{n}\Sigma_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

- Optimisation:

$$argmin_{\hat{Y}}\frac{1}{n}\Sigma_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Machine Learning 101

- Model: $\boxed{Y = f(X)}$ **Machine**

- Objective function (e.g.,): $MSE = \frac{1}{n}\Sigma_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

- Optimisation: $argmin_{\hat{Y}}\frac{1}{n}\Sigma_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

Learning

Classify documents with supervised learning

- Considerations:
    - Feature representation (Bag of words representation or embeddings)
    - Feature selection (remove irrelevant features)
    - Classifier selection
        - E.g., Naive Bayes, SVM, KNN, or ensemble methods

Checking Performance

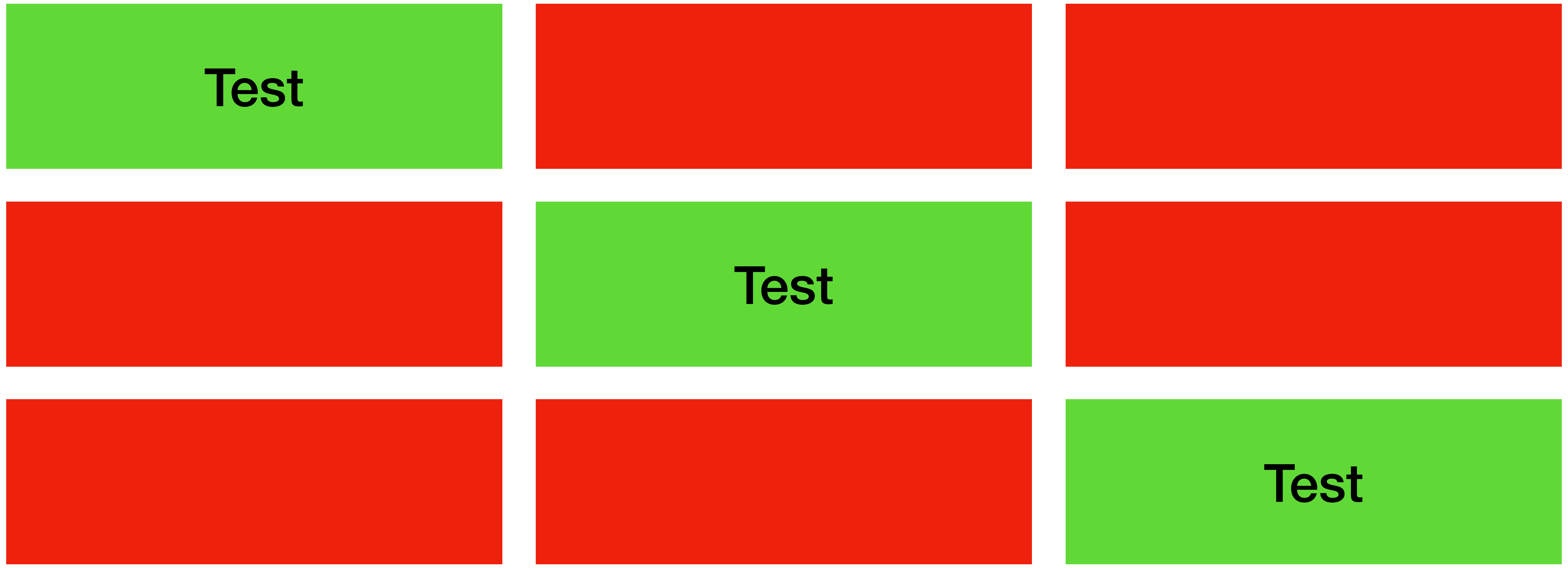
- The fitted model (the trained classifier) is applied to a held-out test set (which is a part of the labeled set but was not used for training the model).

- Considerations:
    - Danger of overfitting (focus on features that work well with training set but do not generalize)
    - Solutions: cross-validation
    - Performance metric (i.e., recall, precision)

Checking performance

- k-fold cross-validation
    - We randomly split the data into k sets (“folds”) of roughly equal size
    - Each set is hold out once as test set, while training on the remaining sets
    - The problem of a lucky split is reduced

K-Fold Cross-Validation



Confusion Matrix

	Actual label	
	Negative	Positive
Negative	True negative	False positive
Positive	False negative	True positive

Precision/Recall

$$Accuracy = \frac{True\ Negative + True\ Positive}{True\ Negative + True\ Positive + False\ Negative + True\ Positive}$$

$$Precision_{positive} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall_{positive} = \frac{True\ Positive}{True\ Positive + False\ Negatives}$$

Dictionary vs. supervised machine learning

- Dictionaries can be applied directly to a new corpus (but validate!)
- Supervised machine learning requires (potentially larger amounts) labeled data
- If the training sample is large enough supervised learning will outperform dictionaries

Additional considerations

- Hyperparameter selection
  - Via systematic comparison of different hyperparameters per algorithm
- Random undersampling (Galar et al., 2011)
- Method to deal with unbalanced classes: use the max. number of positive instances per class and randomly sample the same number of instances of the negative class