CS578 – Interactive and Transparent Machine Learning

Topic: Feature Importance

Mustafa Bilgic

http://www.cs.iit.edu/~mbilgic

https://twitter.com/bilgicm

Transparency

- Which features are important?
- How important is each feature?
- How much does the importance of feature *i* decrease / increase after we add feature *j* to our domain?
- Is feature *i* positively or negatively correlated with the target feature?

Three Main Approaches

- Filter
  - Independent of the classification / regression model
  - E.g., mutual information
- Wrapper
  - Use the underlying model to select a subset of the features
  - E.g., could be done greedily bottom up or top down
- Joint optimization
  - Perform model training and feature selection jointly
  - E.g., L_1-regularized logistic regression

Relevant but different topic

- Feature construction
- Difference
  - Feature selection *selects* a subset of features
  - Feature construction *constructs* a new feature space
- Examples of feature construction
  - Principal Component Analysis (PCA)
  - Factor Analysis (FA)
  - Linear Discriminant Analysis
  - See
    - https://scikit-learn.org/stable/modules/classes.html#module-sklearn.decomposition
    - https://scikit-learn.org/stable/modules/classes.html#module-sklearn.discriminant_analysis

Let's see a few examples

- Now – filter
  - Mutual information
  - Information gain
  - Chi^2
- Later – subset and joint
  - Feature importances in decision trees
  - Feature weights in linear models
  - Regularization (L_1 and L_2)
  - Influence analysis

Mutual Information

$$MI(X,Y) = \sum_{x,y} \hat{P}(x,y) \log \frac{\hat{P}(x,y)}{\hat{P}(x)\hat{P}(y)}$$

Information Gain

$$\text{Entropy} = -\sum_j p_j \log_2 p_j$$

$$InformationGain(X_i) = \text{Entropy before } X_i - \text{ Expected entropy after } X_i$$

Note: information gain can be defined with respect to other 'uncertainty' measures; for example, instead of 'entropy' one can use 'gini' index.

Exercise

- Show that MI and IG are equivalent when IG is defined as reduction in entropy

Chi^2

$$\chi^2(X,Y) = \sum_{x,y} \frac{\left(M[x,y] - M \times \hat{P}(x) \times \hat{P}(y)\right)^2}{M \times \hat{P}(x) \times \hat{P}(y)}$$

Examples

- Synthetic
  - https://github.com/CS578-S19/CS578/tree/master/hugin
- Real datasets
  - https://scikit-learn.org/stable/datasets/index.html

References

- JMLR article
  - http://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf
- Scikit-learn
  - https://scikit-learn.org/stable/modules/classes.html#module-sklearn.feature_selection
- Google Scholar
  - https://scholar.google.com/scholar?q=feature+selection
- Wikipedia
  - https://en.wikipedia.org/wiki/Feature_selection

Your Task

- Read the references listed in the previous slide
- Get familiar with feature selection and ranking approaches on scikit-learn
- Create a synthetic dataset using Hugin and calculate feature importances
- Calculate feature importances for several real datasets
- Note: use Jupyter notebooks wherever possible
- Assignment 1 is coming soon