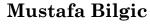
CS578 – INTERACTIVE AND TRANSPARENT MACHINE LEARNING

TOPIC: FEATURE IMPORTANCE





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
 - Chi²
- Later subset and joint
 - Feature importances in decision trees
 - Feature weights in linear models
 - Regularization (L_1 and L_2)
 - Influence analysis

MUTUAL INFORMATION

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

INFORMATION GAIN

Entropy =
$$-\sum_{j} p_{j} \log_{2} p_{j}$$

 $InformationGain(X_i) = Entropy before X_i - Expected entropy after X_i$

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