Feature Selection

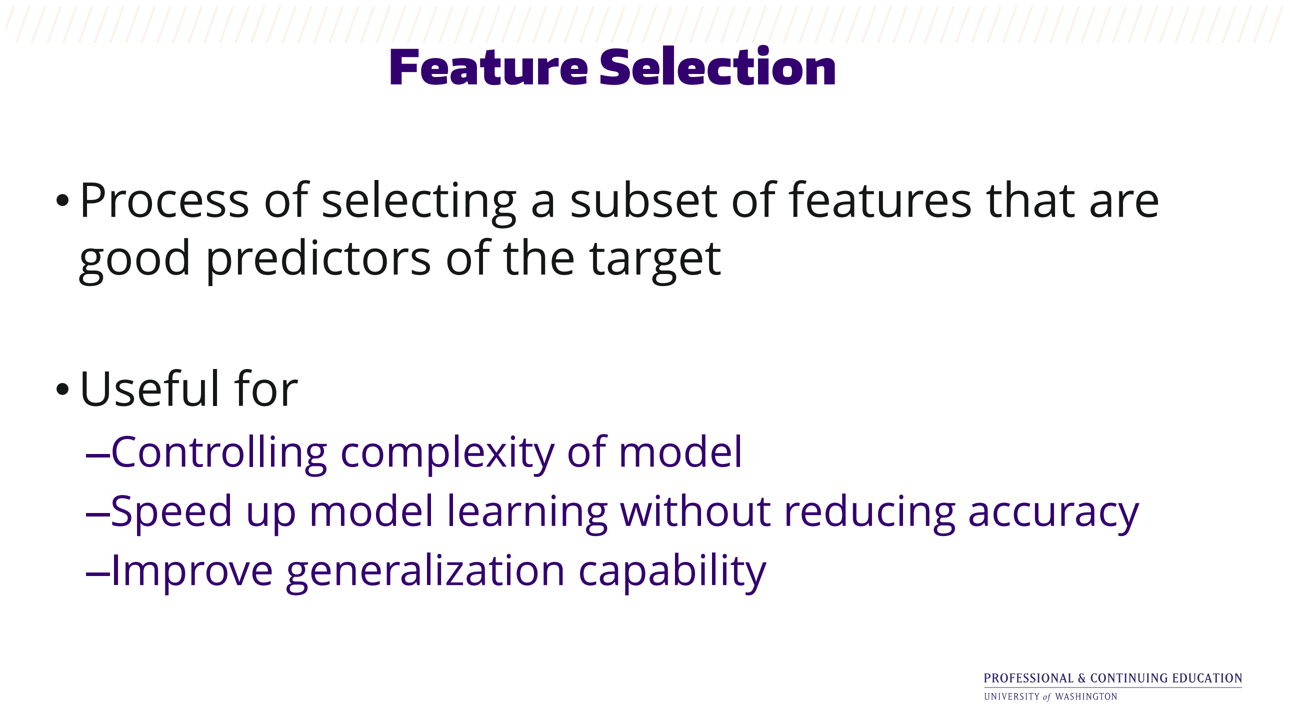
Lesson 6 – Section 3

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Overview

- Why feature selection?
- 3 types of feature selection methods
- Mutual information

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

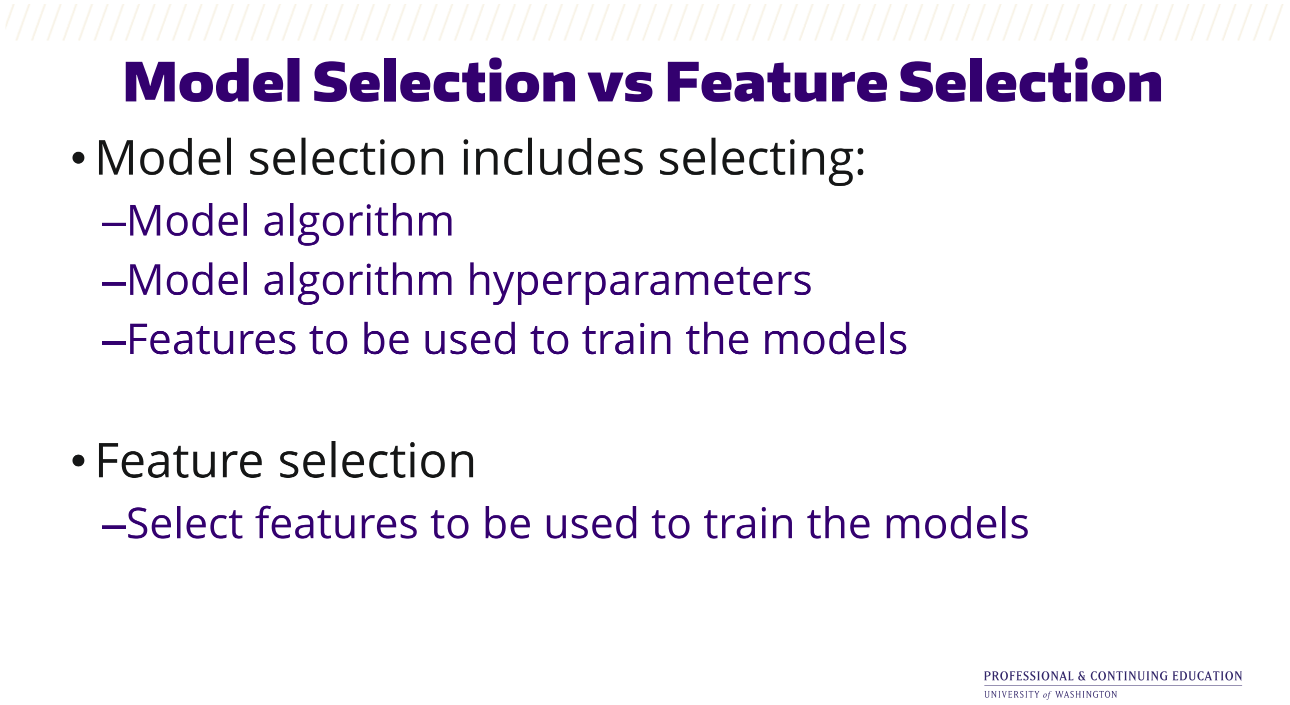


Feature Selection

- Process of selecting a subset of features that are good predictors of the target

- Useful for
  - Controlling complexity of model
  - Speed up model learning without reducing accuracy
  - Improve generalization capability

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY *of* WASHINGTON



Model Selection vs Feature Selection

- Model selection includes selecting:
  - Model algorithm
  - Model algorithm hyperparameters
  - Features to be used to train the models

- Feature selection
  - Select features to be used to train the models

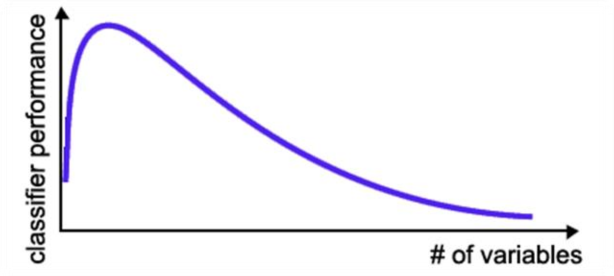
PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY *of* WASHINGTON

Why We Need Feature Selection?

Curse of Dimensionality

- The required number of samples (to achieve the same accuracy) grows exponentially with the number of variables!
- In practice: number of training examples is fixed!

  the classifier's performance will degrade for a large number of features!



In many cases the information lost by discarding variables is made up for by a more accurate mapping/sampling in the lower-dimensional space !

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

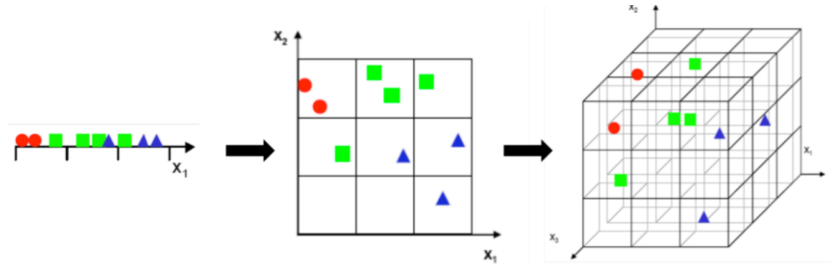
Problems of High-Dimensional Data

- High-dimensional data is often notorious to tackle due to the curse of dimensionality
  - –Increase storage and running time
  - –Overfit the machine learning models
  - –Require more data
- The intrinsic dimension of data may be small
  - –The number of genes responsible for a certain disease

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Curse of Dimensionality – Required Samples

- Data sparsity becomes exponentially worse as feature dimension increases
  - –Conventional distance metrics become ineffective
  - –All points in the high-dimensional space look equally distant



3 samples per region | 1 sample per region | 1/3 sample per region

http://nikhilbuduma.com/2015/03/10/the-curse-of-dimensionality/

Feature Selection, 3 types of methods (1)

Filter Methods, select a subset of features before training a model, e.g.

- –Correlation with target,
- –Mutual Information between feature and target

•*Simple to implement, and have reasonable performance*

Feature Selection, 3 types of methods (2)

Wrapper Methods, search combination of feature space by training and evaluating model using a subset of features, e.g.

- –Forward, backward, step-wise feature selection,
- –Genetic algorithms.

- •*Computationally expensive and prone to over-fitting*

Feature Selection, 3 types of methods (3)

Embedded Methods, feature subset is chosen as part of model training, e.g.

- –LASSO (L-1) regression, Regularized **decision trees, random forests**

- •*Typically robust to over-fitting, but has hyper parameters that will need to be fit using a validation data*

Filter-Based Feature Selection

Filter-based Feature Selection

- Correlation with target variable
  - A good starting point
  - If Y is categorical variable (classification):
    - Use chi-square test to decide the correlation between each categorical X variable and Y variable
    - Use ANOVA test to decide the correlation between each numerical X variable and Y variable
  - If Y is continuous variable (regression):
    - Use ANOVA test to decide the correlation between each categorical X variable and Y variable
    - Use correlation between each numerical X variable and Y variable

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY *of* WASHINGTON

Filter-based Feature Selection

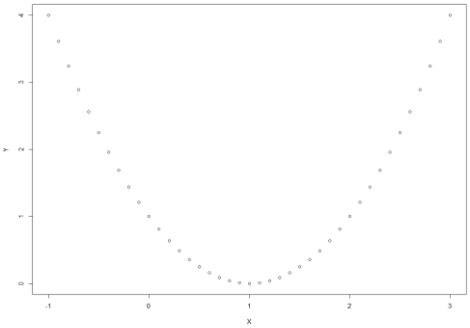
Alert:

If x1 and x2 are highly correlated, and x1 and Y are highly correlated, both x1 and x2 will be selected based on correlation with Y.

- Strong correlations in X will bring some challenge for some machine learning models, such as linear regression model.

Is Correlation Always a Good Choice?

- It makes sense for linear regression (logistic regression) model.
  - Since linear regression model only looks at linear relationship
- Does not make sense for nonlinear models such as tree-based models
- Cannot capture nonlinear relationship between X and Y



Mutual Information

- Captures Statistical Dependency between Two Variables $\Pr(X,Y) = \Pr(X) \times \Pr(Y)$
  - –If two variables are statistically independent

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$$

$$\hat{f}(x) = \frac{1}{Nh\sqrt{2\pi}} \sum_{i=1}^{N} \exp\left(\frac{-(x-x_i)^2}{2h^2}\right).$$

  - –Estimate Pr(X) from observations by using a kernel function

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Summary

Feature Selection to avoid high-dimension sparse data

>Filter Methods

–Subset of data before splitting based on correlation or mutual information

>Mutual Information

–Captures the statistical dependency between 2 variables