Feature Engineering

Papangkorn Inkeaw, Ph.D.

Dimensionality Reduction

Chapter 6 (Part I) - Feature Selection

Dimensionality Reduction

High dimensional data:

- High computational cost
- Over-fitting when learning a model or requiring large amounts of training data
- Highly correlated and hence redundant.
- Distances between data points can become equidistant.

Importance of Dimensionality Reduction

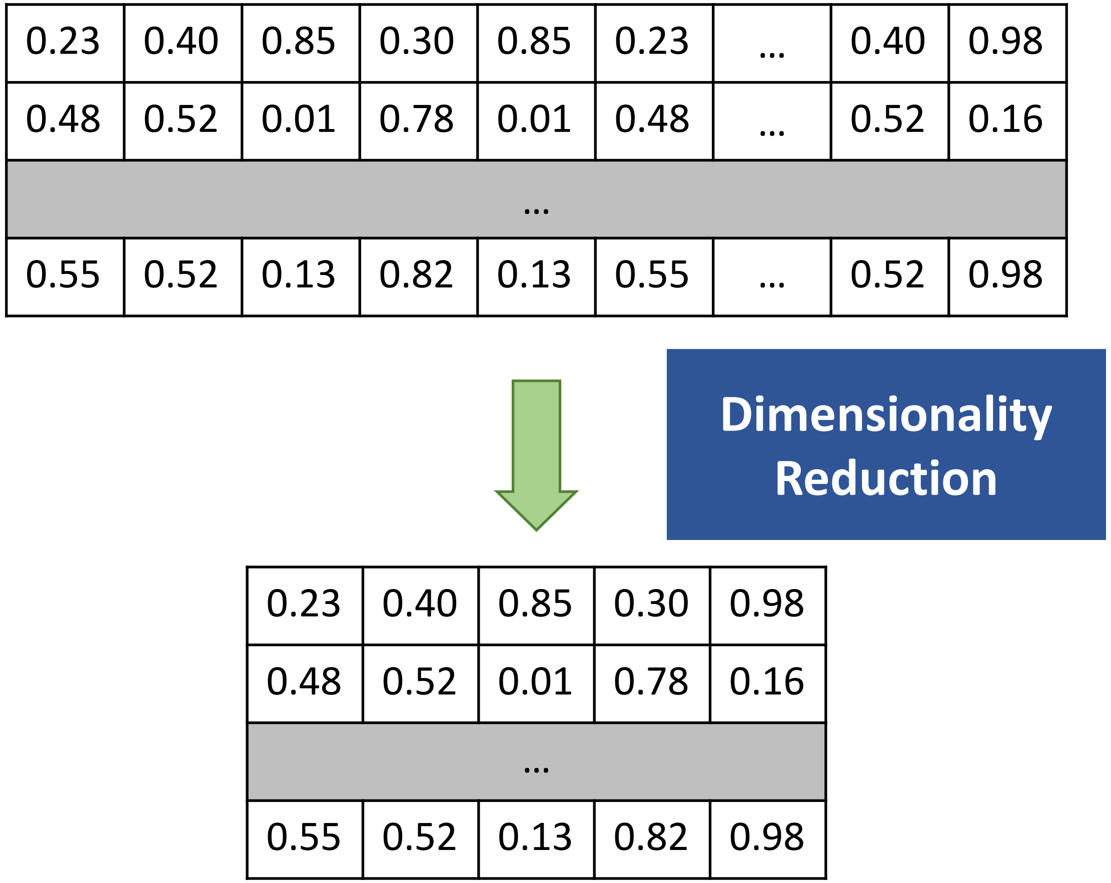
- Reduce computational time and space
- Remove multi-collinearity and irrelevant features (can improv predictive performance)
- Avoids the curse of dimensionality

Dimensionality Reduction

Goal of Dimensionality Reduction:

Reduce the number of predictors as far as possible without compromising predictive performance.

Dimensionality Reduction



0.23	0.40	0.85	0.30	0.85	0.23	…	0.40	0.98
0.48	0.52	0.01	0.78	0.01	0.48	…	0.52	0.16
…								
0.55	0.52	0.13	0.82	0.13	0.55	…	0.52	0.98

Dimensionality Reduction

0.23	0.40	0.85	0.30	0.98
0.48	0.52	0.01	0.78	0.16
…				
0.55	0.52	0.13	0.82	0.98

Feature Selection

Find a subset of the input variables.

Feature Projection

Transforms the data in the high-dimensional space to a space of fewer dimensions.

Feature Selection

Categories of Feature Selection Methods:

1. **Filter methods**
   - Use a proxy measure to score a feature subset.
2. **Wrapper methods**:
   - Use a predictive model to score feature subsets.
3. **Embedded methods**
   - Perform feature selection as part of the model construction process.

Filter Methods

- Popularly Operate over only one feature and their value distribution, together with the target class.
- Use a feature utility metric to measure the predictive ability.
- Feature Utility Metrics:
    - Chi-square *(For a categorical feature and a categorical target)*
    - Pearson Correlation Coefficient (*For a categorical feature and a continuous target)*
    - Mutual Information *(For a categorical feature and a categorical target)*
    - Gini Impurity *(For a categorical feature and a categorical target)*
    - ANOVA test *(For a categorical feature and a continuous target)*

Filter Methods

Chi-Square

- Measure the correlation between one categorical feature and the target class.
- A statistic computed from the confusion table of count
- It rejects the null hypothesis that the feature chosen and the target class happen at random.
- Chi-square is the difference between the observed values and expected values:

$$\chi^2 = \sum_{i,j} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

Filter Methods

Chi-Square

Contingency table

Class/ Gender	Yes	No	Total
Male	$O_{Male,Yes}$ 38	$O_{Male,No}$ 178	216 (p=216/400=0.54)
Female	$O_{Female,Yes}$ 44	$O_{Female,No}$ 140	184 (p=82/400=0.46)
Total	82 (p=82/400=0.205)	318 (p=318/400=0.795)	n=400

Degrees of freedom = (r-1)*(c-1)=(2-1)*(2-1)=1

$$\chi^2 = \frac{(38-44.28)^2}{44.28} + \frac{(178-171.72)^2}{181.72} + \frac{(44-37.72)^2}{37.72} + \frac{(140-146.28)^2}{146.28}$$

= 2.4354

Expected values:

$E_{Male,\ Yes}$ = n*(p(Yes)*p(Male))
= 400*((82/400)*(216/400))
= 400*(0.205*0.54)
=400*0.117
=44.28

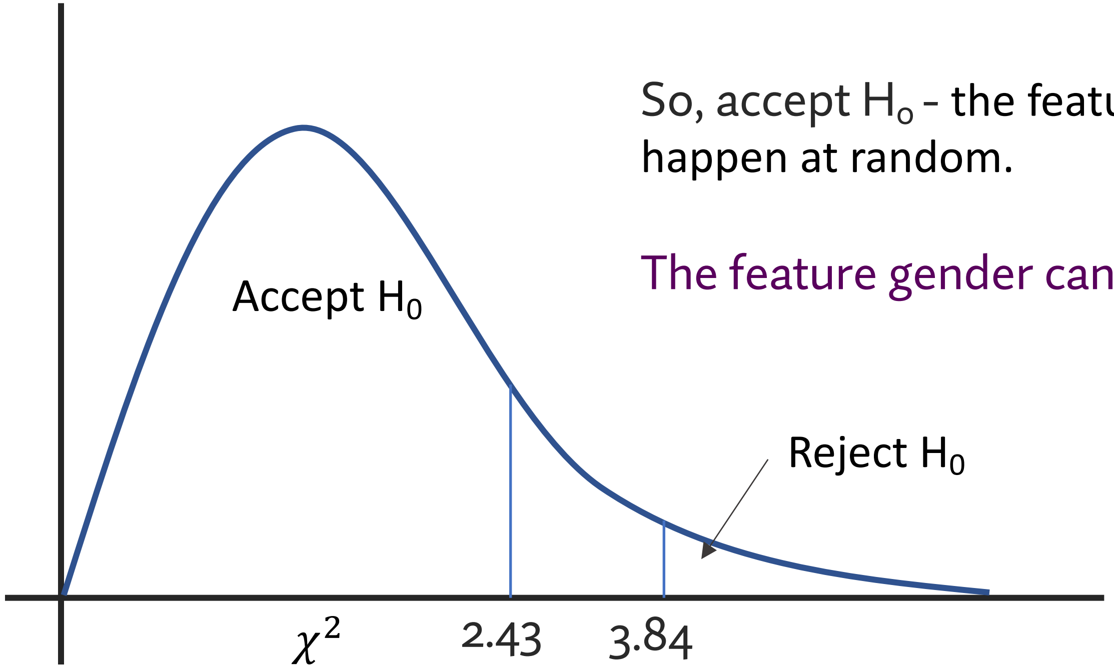
$E_{Male,\ No}$ = n*(p(No)*p(Male))
= 400*(0.795*0.54)
=171.72

$E_{Female,\ Yes}$ = n*(p(Yes)*p(Female))
= 400*(0.205*0.46)
=37.72

$E_{Female,\ No}$ = n*(p(No)*p(Female))
= 400*(0.795*0.46)
=146.28

Filter Methods

Chi-Square



So, accept H_o - the feature chosen and the target class happen at random.

The feature gender cannot be selected for training model.

At degrees of freedom =1 and confidence = 95%, the Chi-Square value is 3.84.

Filter Methods

Pearson Correlation Coefficient

- Captures a goodness of linear fit on individual features.
- It is defined as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2(y_i - \bar{y})^2}}$$

- The coefficient ranges from 1.0 (perfect correlation) to –1.0 (perfect anticorrelation).
- Values closer to zero indicate no relation.
- Feature selection uses the square of the coefficient r^2.

Filter Methods

Mutual Information

- The number of bits of information that is known about a second random variable if we know a first random variable.
- It is related to information theory's *entropy* of a random variable, the amount of information held by the variable.
- For *m* categories, the entropy can be computed by

$$H(X) = \sum_{v \in possible\ values\ of\ X} -P(x = v) \log_2 P(x = v)$$

Filter Methods

Mutual Information

- Mutual Information is the entropy of one of the variables minus the conditional entropy ofone variable given the other:

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

- Mutual information is always larger than or equal to zero,
  - The greater the relationship between the two variables.
  - If the calculated result is zero, then the variables are independent.

Filter Methods

Gini Impurity

- It splits the instances into easier to solve sub-problems.
- Given the split induced in the instances by a feature X, Gigi impurity can be computed by

$$Gini - Impurity(X) = \sum_{v}^{X} \frac{count(x = v)}{N} \left(1 - \sum_{c}^{target} \left(\frac{count(x = v \; and \; t = c)}{count(x = v)} \right)^2 \right)$$

- Features with lower Gini-Impurity should be preferred.

Filter Methods

Gini Impurity

Contingency table			
Class/ Gender	Yes	No	Total
Sick	1	2	3
Not Sick	3	2	5
Total	4	4	n=8

$$Gini - Impurity(Gender) = \sum_{v \in \{Sick, Not-Sick\}} count(x = v)\left(1 - \sum_{c \in \{Yes, No\}} \left(\frac{count(x = v\ and\ t = c)}{count(x = v)}\right)^2\right)$$

$$= \left(\frac{3}{8}\right)\left(1 - \left(\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right)\right) + \left(\frac{5}{8}\right)\left(1 - \left(\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right)\right)$$

$$= 0.47$$

Filter Methods

ANOVA

- It tests whether there is a difference between groups, by comparing their means and their variances.
  - The mean and variances are from values of the target
  - produced by grouping the different values of the categorical feature.
- The expectation is that a feature where the target value is different for different categorical values will be an useful regressor.
- This metric returns an F-statistic that indicates whether the groups are similar or different.

Filter Methods

ANOVA

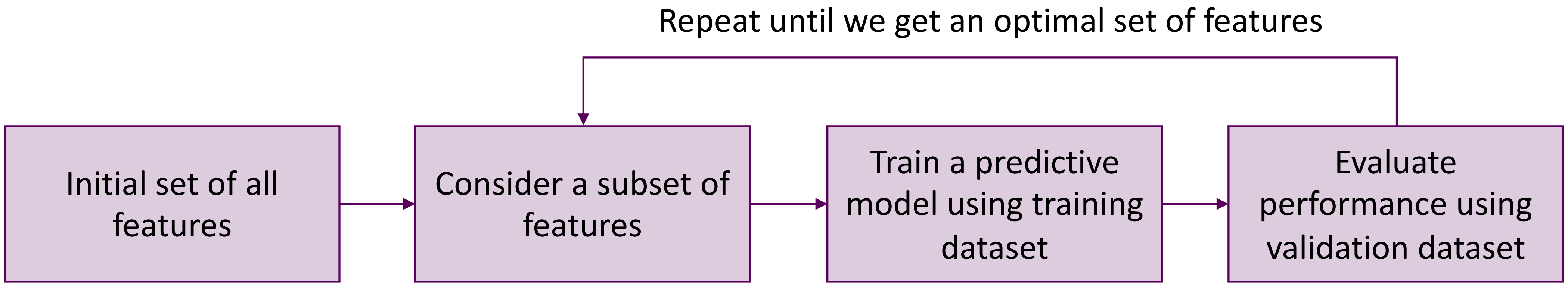
Null hypothesis: No real difference exists between the tested groups.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F-statistic
Within	$SS_W = \sum_{j=1}^{k} \sum_{i=1}^{n} (X_{ij} - \bar{X}_j)^2$	$n - k$	$MS_W = \frac{SS_W}{n - k}$	$F = \frac{MS_B}{MS_W}$
Between	$SS_B = \sum_{j=1}^{k} n_j (\bar{X}_j - \bar{X})^2$	$k - 1$	$MS_B = \frac{SS_B}{k - 1}$	
total	$SS_T = \sum_{j=1}^{k} \sum_{i=1}^{n} (X_{ij} - \bar{X})^2$			

k is the number of groups
n is the numbre of samples

Wrapper Methods

- User a specific predictive model to evaluate possible subsets of features.
- Perform a held-out set or cross-validation.
- Approaches:
  - Greedy methods
  - Global search methods


Repeat until we get an optimal set of features
Initial set of all features
Consider a subset of features
Train a predictive model using training dataset
Evaluate performance using validation dataset

Wrapper Methods

Forward Feature Selection

1. Evaluate each individual feature
2. Add the feature with the highest score into the selecting set and add the others into the remaining set.
3. Repeat
    1. For each feature *f* in the remaining set
        1. Evaluate the selected feature(s) together with the feature *f*
    2. Add the feature that can improve predictive performance with the highest score into the selecting set, and remove it from the remaining set
4. Until an addition of a new variable does not improve the performance.

Wrapper Methods

Backward Feature Selection

1. Evaluate full set of features and add all features into the selecting set
2. Repeat
    1. For each feature *f* in the selecting set
        1. Evaluate the selected feature(s) without with the feature *f*
    2. Removes the least significant feature which improves the performance of the model.
3. Util no improvement is observed on the removal of features.

Wrapper Methods

Recursive Feature Elimination

- Given an external estimator that assigns weights to features.
- Select features by recursively considering smaller and smaller sets of features.
  1. The estimator is trained on the initial set of features.
  2. The importance of each feature is obtained either through a coefficient attribute or through a feature importance attribute.
  3. The least important features are pruned from the current set of features.
  4. That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached.

Embedded Methods

- Integrate feature selection as a part of the learning algorithm.
- A learning algorithm performs feature selection and classification/regression at the same time.
- The most common embedded techniques are the tree algorithms:
  - Decision Tree
  - Random Forest
- Other Embedded Methods are:
  - LASSO with the L1 penalty

Embedded Methods

Decision Tree

Construct a decision tree

STEP 1: If the all values of target variable in D are the same or the stopping criteria is met,
construct a leaf node which produce the most target values as the prediction.

STEP 2: For each variable X_i in D, calculate the information gain (separate the dataset D based
on the variable X_i with a cutoff)

STEP 3: Given X* is the variable that its information gain is the highest

STEP 4: Construct a decision node T on the variable X*

Feature selection task

STEP 5: Separate the dataset D into subsets D_v

STEP 6: For each subset D_v

STEP 6.1: Construct a decision tree T_v on the subset

STEP 6.2: Construct a branch that link the decision node T and T_v

Embedded Methods

Random Forest

- Ensemble method: Use many decision trees

Construct a random forest

For b=1,…,B
STEP 1: Sample, with replacement, n training data from the dataset D
STEP 2: Construct a decision T_b using the sampled data

Feature selection is embeded in decision tree construction process

Prediction phase

STEP 1: Given an unseen sample x, perform each decision tree in the forest.
STEP 2: Aggregate B predictions by averaging for regression or majority-voting for classification

Embedded Methods

Regularization

- When we train some predictive models f(x), such as linear regression, SVM and ANN, we need to define a loss function L
- Regularization is a technique used for tuning models by adding an additional penalty term in the loss function.

$$\mathcal{L} = E(y, \hat{y}) + \lambda R(w)$$

where *E* is the error function, *R* is a regularization function and λ determine how big of an effect of regularization.

- Most common regularizations:
  - L1 – LASSO (appropriate for feature selection)
  - L2 – Ridge

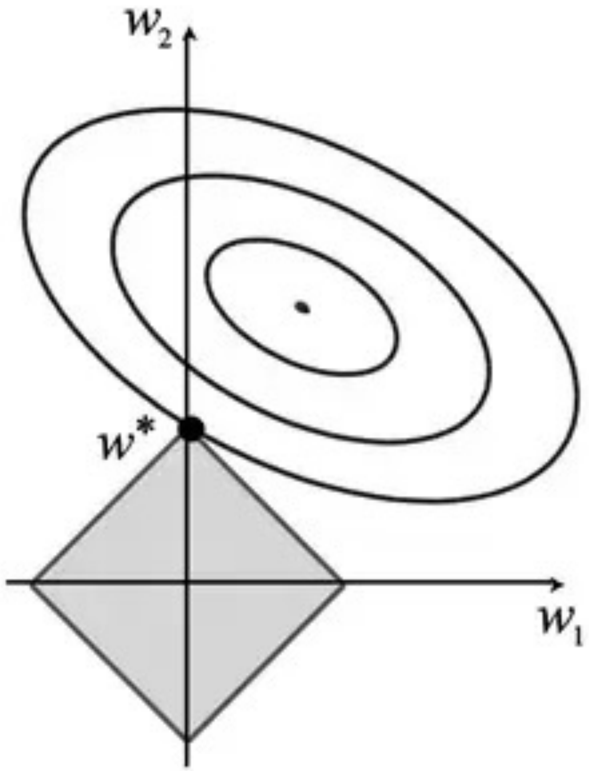
Embedded Methods

LASSO – L1 Regularization

- The sum of the absolute values of all weights in the model.

$$\mathcal{L} = E(y, \hat{y}) + \lambda \sum_{w} |w|$$

- Can result in sparse models with few coefficients
  - Some coefficients, in fact, can become zero and can be eliminated from the model. Feature Selection



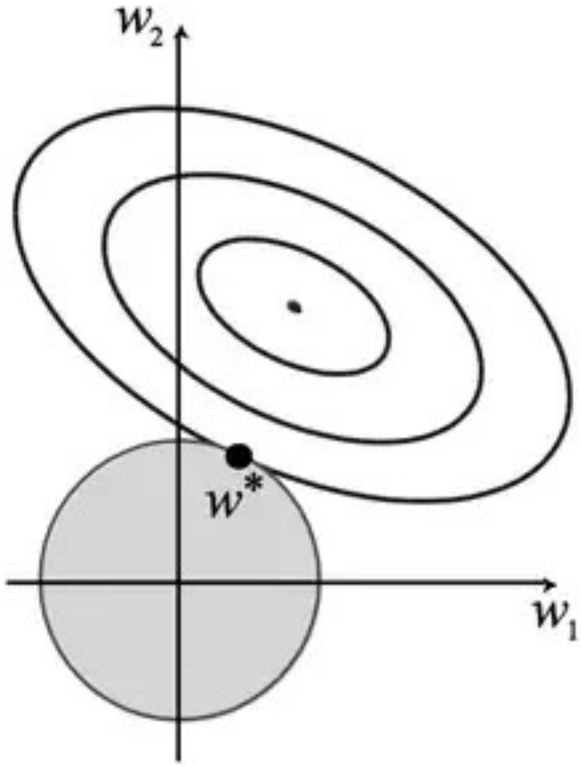
Embedded Methods

Ridge – L2 Regularization

- Adds “squared magnitude of the coefficient” as penalty term to the loss function.

$$\mathcal{L} = E(y, \hat{y}) + \lambda \sum_{w} w^2$$

- It forces the weights to be small but does not make them zero and does not give the sparse solution.
- This technique works very well to avoid over-fitting issue.
- Not specific feature selection mechanism



References & Study Resources

- Max Kuhn and Kjell Johnson. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press.
- Pablo Duboue. (2020). *The Art of Feature Engineering: Essentials for Machine Learning*. Cambridge University Press.
- Mohammed J. Zaki and Wagner Meira JR. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithm*. Cambridge University Press.
- Christopher M. Bishop. (2006). *Pattern Recognition and Machine Learning*. Springer.