

Introduction to Statistical Learning and Applications

Model Selection, Cross Validation and Support Vector Classifier

Razan MHANNA

STATIFY team, Inria centre at the University Grenoble Alpes
Team 5, Grenoble Institute of Neurosciences GIN

25/02/2025



1 Model Selection

2 Cross-validation

- K-Fold Cross-Validation
- Leave-One-Out Cross-Validation (LOOCV)
- Stratified Cross-Validation

3 Support Vector Classifier

- 1 Model Selection
 - Subset Selection
- 2 Cross-validation
 - K-Fold Cross-Validation
 - Leave-One-Out Cross-Validation (LOOCV)
 - Stratified Cross-Validation
- 3 Support Vector Classifier

Model selection

Three classes of methods:

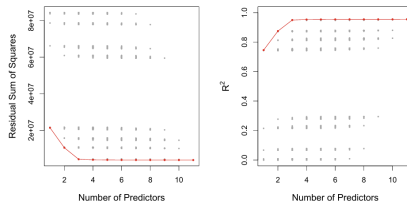
- Subset Selection by taking a subset of the p predictors, then fit a model using least squares on the reduced set of variables.
- Shrinkage: All p predictors are involved, but the estimated coefficients are shrunk towards zero relative to the least squares estimates. The shrinkage (also known as regularization) has the effect of reducing variance and can also perform variable selection.
- Dimension Reduction: Project the p predictors into a M -dimensional subspace, where $M < p$. This is achieved by computing M different linear combinations, or projections, of the variables. Then these M projections are used as predictors to fit a linear regression model by least squares.

- 1 Model Selection
 - Subset Selection
- 2 Cross-validation
 - K-Fold Cross-Validation
 - Leave-One-Out Cross-Validation (LOOCV)
 - Stratified Cross-Validation
- 3 Support Vector Classifier

Best Subset Selection

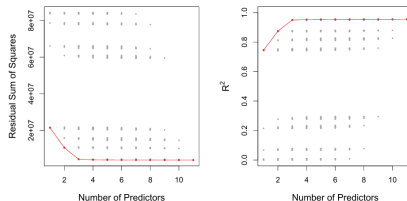
- ① Let \mathcal{M}_0 denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
- ② For $k = 1, 2, \dots, p$:
 - ① Fit all $\binom{p}{k}$ models that contain exactly k predictors.
 - ② Pick the best among these $\binom{p}{k}$ models, and call it \mathcal{M}_k . Here, *best* is defined as having the smallest RSS, or equivalently the largest R^2 .
- ③ Select a single best model from among $\mathcal{M}_0, \dots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2 .

Example dataset



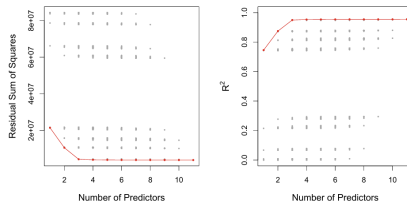
- Each point on the plot represents a least squares regression model built from a different subset of predictors in the Credit data.
- In total, 11 variables are considered (10 predictors plus two dummy variables for the three-level qualitative variable $\hat{a}region\hat{a}$).
- The graph plots the RSS and R^2 as functions of the number of predictors, with red curves linking the best models at each size.
- The results show that while model performance improves with more predictors, beyond a three-variable model, additional predictors yield only minimal gains.

Example dataset



- Each point on the plot represents a least squares regression model built from a different subset of predictors in the Credit data.
- In total, 11 variables are considered (10 predictors plus two dummy variables for the three-level qualitative variable $\hat{a}region\hat{a}$).
- The graph plots the RSS and R^2 as functions of the number of predictors, with red curves linking the best models at each size.
- The results show that while model performance improves with more predictors, beyond a three-variable model, additional predictors yield only minimal gains.

Example dataset



- Each point on the plot represents a least squares regression model built from a different subset of predictors in the Credit data.
- In total, 11 variables are considered (10 predictors plus two dummy variables for the three-level qualitative variable $\hat{a}region\hat{a}$).
- The graph plots the RSS and R^2 as functions of the number of predictors, with red curves linking the best models at each size.
- The results show that while model performance improves with more predictors, beyond a three-variable model, additional predictors yield only minimal gains.

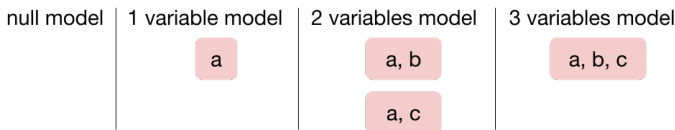
Limitations of Best Subset Selection

- For computational reasons, best subset selection cannot be applied with very large p . *Why not?*
- Best subset selection may also suffer from statistical problems when p is large: larger the search space, the higher the chance of finding models that look good on the training data, even though they might not have any predictive power on future data.
- Thus an enormous search space can lead to *overfitting* and high variance of the coefficient estimates.
- For both of these reasons, *stepwise* methods, which explore a far more restricted set of models, are attractive alternatives to best subset selection.

Forward Stepwise Selection

- Forward stepwise selection begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model.
- In particular, at each step the variable that gives the greatest *additional* improvement to the fit is added to the model.

Forward stepwise selection



Forward Stepwise Selection

Forward Stepwise Selection

- 1 Let \mathcal{M}_0 denote the *null* model, which contains no predictors.
- 2 For $k = 0, \dots, p-1$:
 - 1 Consider all $p-k$ models that augment the predictors in \mathcal{M}_k with one additional predictor.
 - 2 Choose the *best* among these $p-k$ models, and call it \mathcal{M}_{k+1} . Here *best* is defined as having smallest RSS or highest R^2 .
- 3 Select a single best model from among $\mathcal{M}_0, \dots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2 .

More on Forward Stepwise Selection

- Computational advantage over best subset selection is clear.
- It is not guaranteed to find the best possible model out of all 2^p models containing subsets of the p predictors. *Why not? Give an example.*

Concerns about model selection

- Given several candidate models, we can use **cross validation** to compare their predictive performance. But does it always make sense to choose the model that optimizes estimated predicted performance?
 - 1 If the only goal is prediction on data similar to what was observed, it can be better to average across models rather than to choose just one.
 - 2 If the goal is prediction on a new set of data, we should reweight the predictive errors accordingly to account for differences between sample and population,
 - 3 Rather than choosing among or averaging over an existing set of models, a better choice can be to perform continuous model expansion, constructing a larger model that encompasses the earlier models as special cases.

1 Model Selection

2 Cross-validation

- Definition
- Types of cross-validation
 - K-Fold Cross-Validation
 - Leave-One-Out Cross-Validation (LOOCV)
 - Stratified Cross-Validation

3 Support Vector Classifier

1 Model Selection

2 Cross-validation

- **Definition**

- Types of cross-validation

- K-Fold Cross-Validation
- Leave-One-Out Cross-Validation (LOOCV)
- Stratified Cross-Validation

3 Support Vector Classifier

Cross-Validation

Definition

Cross-validation is a statistical method used to estimate the performance (or accuracy) of predictive statistical models.

- It is used to protect against overfitting in a predictive model, particularly in a case where the amount of data may be limited.
- In cross-validation, you make a fixed number of folds (or partitions) of the data, run the analysis on each fold, and then average the overall error estimate.

[1] Jun Shao.

Linear model selection by cross-validation.

Journal of the American Statistical Association, 88:486–494, 1993.

1 Model Selection

2 Cross-validation

- Definition

- Types of cross-validation

- K-Fold Cross-Validation
- Leave-One-Out Cross-Validation (LOOCV)
- Stratified Cross-Validation

3 Support Vector Classifier

Types of Cross Validation Methods

There are different types of cross-validation methods, and they could be classified into two broad categories:

- **Non-exhaustive**

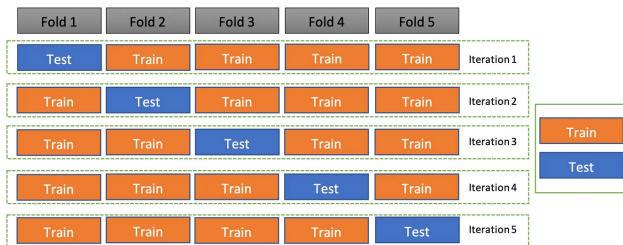
- Holdout
- K-Fold
- Stratified K-Fold

- **Exhaustive Methods**






- Leave-P-Out cross-validation
- Leave-1-Out cross-validation

K-Fold Cross-Validation






- 1 Randomly split your dataset into k subsets (folds).
- 2 For each fold:
 - Train the model using $k - 1$ folds.
 - Test the model on the remaining fold.
- 3 Repeat until each fold has been used as a test set.
- 4 Compute the average accuracy across all folds.








Advantages of k-Fold Cross-Validation

- Low bias and variance 
- Better estimation of model performance 
- Maximizes data utilization 
- Reduces overfitting
- Supports hyperparameter tuning 
- Provides insights into model behavior 
- Useful for small datasets
- Suitable for imbalanced datasets






Advantages of k-Fold Cross-Validation

- Low bias and variance 
- Better estimation of model performance 
- Maximizes data utilization 
- Reduces overfitting
- Supports hyperparameter tuning 
- Provides insights into model behavior 
- Useful for small datasets
- Suitable for imbalanced datasets






Advantages of k-Fold Cross-Validation

- Low bias and variance 
- Better estimation of model performance 
- Maximizes data utilization 
- Reduces overfitting
- Supports hyperparameter tuning 
- Provides insights into model behavior 
- Useful for small datasets
- Suitable for imbalanced datasets






Limitations

- High computational cost 
- Not suitable for very large datasets 
- Time-consuming for complex models 
- Potential information leakage 
- Not suitable for time-series data
- Results can vary based on k 

Limitations

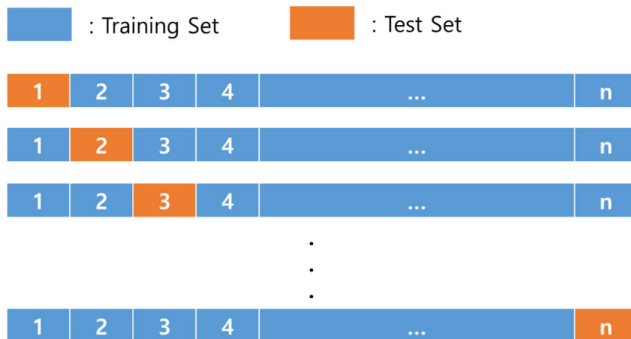
- High computational cost 
- Not suitable for very large datasets 
- Time-consuming for complex models 
- Potential information leakage 
- Not suitable for time-series data
- Results can vary based on k 

Limitations

- High computational cost 
- Not suitable for very large datasets 
- Time-consuming for complex models 
- Potential information leakage 
- Not suitable for time-series data
- Results can vary based on k 

Leave-One-Out Cross-Validation (LOOCV)

- LOOCV is a cross-validation technique where a single data point is used as the validation set, and the rest of the data is used for training.
- This process is repeated for each data point, ensuring that each observation is used once as the validation data.

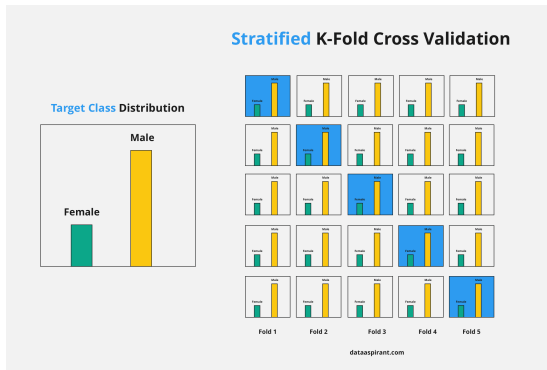


Advantages and Limitations

Advantages	Limitations
—Utilizes all available data for training and validation.	—Computationally expensive for large datasets.
—Provides an unbiased estimate of model performance.	—High variance in model performance estimation.
—Low bias as each model is trained on almost all data points.	—Not suitable for large sample sizes due to limited improvement over other methods.

Stratified Cross-Validation

Stratified cross-validation is a variation of k -fold cross-validation where each fold or partition of the data preserves the proportion of samples from each class or category. This ensures that the distribution of classes is consistent across all folds, making it particularly useful for classification tasks with imbalanced class distributions.



Advantages

- Preserves Class Distribution where each fold contains a representative sample of all classes,
- By preserving the class distribution in each fold, stratified cross-validation provides a more accurate estimate of model performance, leading to better generalization to unseen data.
- Helps to reduce bias in model evaluation,
- Improved Model Evaluation: Provides a more reliable assessment of model performance,

Advantages

- Preserves Class Distribution where each fold contains a representative sample of all classes,
- By preserving the class distribution in each fold, stratified cross-validation provides a more accurate estimate of model performance, leading to better generalization to unseen data.
- Helps to reduce bias in model evaluation,
- Improved Model Evaluation: Provides a more reliable assessment of model performance,

Advantages

- Preserves Class Distribution where each fold contains a representative sample of all classes,
- By preserving the class distribution in each fold, stratified cross-validation provides a more accurate estimate of model performance, leading to better generalization to unseen data.
- Helps to reduce bias in model evaluation,
- Improved Model Evaluation: Provides a more reliable assessment of model performance,

Disadvantages

- Increased Computational Cost compared to simple random sampling, as it involves the extra step of stratifying the data based on class labels.
- Complex Implementation, particularly when dealing with large and multi-class datasets.
- Dependency on Class Labels: Relies on accurate class labels for stratification, which may be challenging to obtain in certain real-world datasets or may introduce biases if the labels are noisy or inaccurate.
- In some cases, stratified cross-validation may lead to overfitting, particularly if the class distribution in the dataset is highly skewed or if the number of samples per class is very small.
- While highly beneficial for classification tasks, stratified cross-validation is not directly applicable to regression problems, where the target variable is continuous rather than categorical.

Disadvantages

- Increased Computational Cost compared to simple random sampling, as it involves the extra step of stratifying the data based on class labels.
- Complex Implementation, particularly when dealing with large and multi-class datasets.
- Dependency on Class Labels: Relies on accurate class labels for stratification, which may be challenging to obtain in certain real-world datasets or may introduce biases if the labels are noisy or inaccurate.
- In some cases, stratified cross-validation may lead to overfitting, particularly if the class distribution in the dataset is highly skewed or if the number of samples per class is very small.
- While highly beneficial for classification tasks, stratified cross-validation is not directly applicable to regression problems, where the target variable is continuous rather than categorical.

Disadvantages

- Increased Computational Cost compared to simple random sampling, as it involves the extra step of stratifying the data based on class labels.
- Complex Implementation, particularly when dealing with large and multi-class datasets.
- Dependency on Class Labels: Relies on accurate class labels for stratification, which may be challenging to obtain in certain real-world datasets or may introduce biases if the labels are noisy or inaccurate.
- In some cases, stratified cross-validation may lead to overfitting, particularly if the class distribution in the dataset is highly skewed or if the number of samples per class is very small.
- While highly beneficial for classification tasks, stratified cross-validation is not directly applicable to regression problems, where the target variable is continuous rather than categorical.

Disadvantages

- Increased Computational Cost compared to simple random sampling, as it involves the extra step of stratifying the data based on class labels.
- Complex Implementation, particularly when dealing with large and multi-class datasets.
- Dependency on Class Labels: Relies on accurate class labels for stratification, which may be challenging to obtain in certain real-world datasets or may introduce biases if the labels are noisy or inaccurate.
- In some cases, stratified cross-validation may lead to overfitting, particularly if the class distribution in the dataset is highly skewed or if the number of samples per class is very small.
- While highly beneficial for classification tasks, stratified cross-validation is not directly applicable to regression problems, where the target variable is continuous rather than categorical.

Disadvantages

- Increased Computational Cost compared to simple random sampling, as it involves the extra step of stratifying the data based on class labels.
- Complex Implementation, particularly when dealing with large and multi-class datasets.
- Dependency on Class Labels: Relies on accurate class labels for stratification, which may be challenging to obtain in certain real-world datasets or may introduce biases if the labels are noisy or inaccurate.
- In some cases, stratified cross-validation may lead to overfitting, particularly if the class distribution in the dataset is highly skewed or if the number of samples per class is very small.
- While highly beneficial for classification tasks, stratified cross-validation is not directly applicable to regression problems, where the target variable is continuous rather than categorical.

1 Model Selection

2 Cross-validation

- K-Fold Cross-Validation
- Leave-One-Out Cross-Validation (LOOCV)
- Stratified Cross-Validation

3 Support Vector Classifier