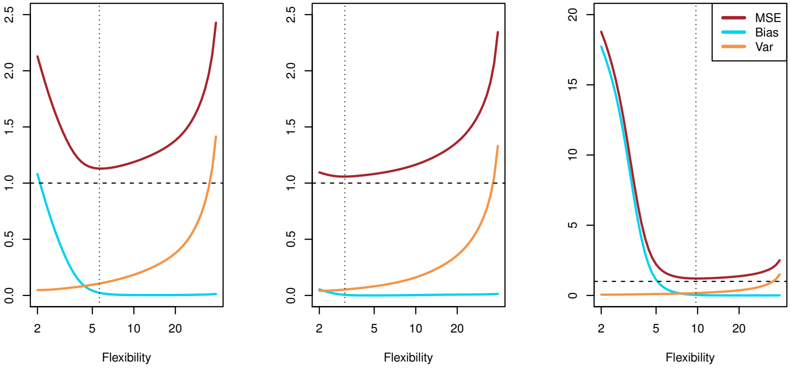
Model Assessment and Hyperparameter Tuning

Johanni Brea

Introduction à l'apprentissage automatique

GYMINF 2021

Which Model Is Best?

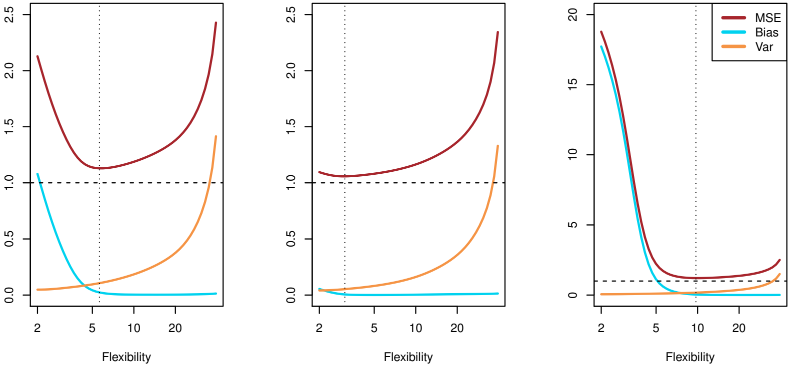


The best model has smallest test MSE†.

† Here the test MSE is exactly computed.

Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

Which Model Is Best?



The best model has smallest test MSE†.

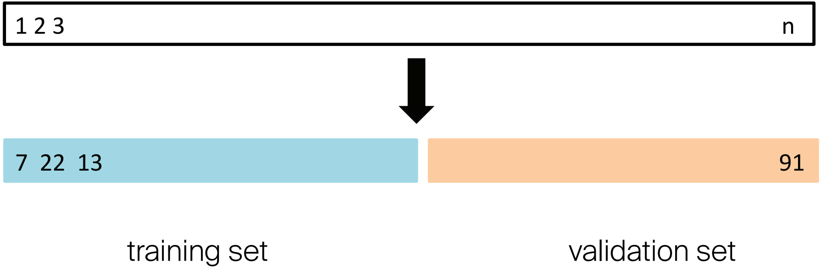
What if we do not know the true test error?

† Here the test MSE is exactly computed.

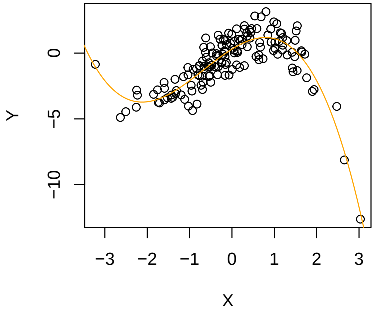
Some of the figures in this presentation are taken from “An Introduction to Statistical Learning, with applications in R” (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

Table of Contents

The Validation Set Approach

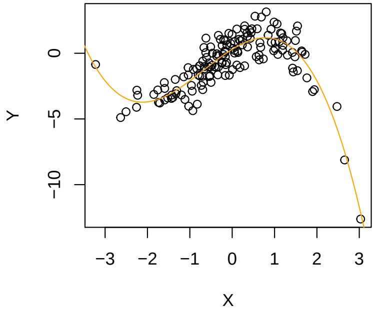


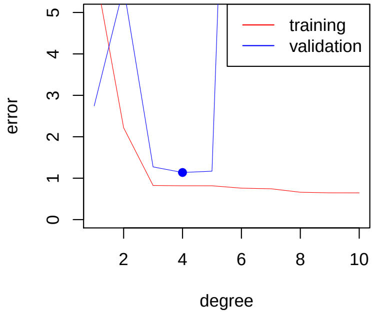
Validation Set Approach Applied to Artificial Data



$$Y = 0.3 + 2X - 0.8X^2 - 0.4X^3 + \epsilon$$

Validation Set Approach Applied to Artificial Data





$$Y = 0.3 + 2X - 0.8X^2 - 0.4X^3 + \epsilon$$

polynomial fits with different degrees d

optimal $d = 4$

Training, Validation and Test Set

- **Training Set**: Subset of the full data used to find the parameters.
- **Validation Set**: Held-out subset of the full data used for model selection, i.e. finding the hyper-parameters.
- **Test Set**: Held-out subset of the data to estimate the test error of the best model.

Training, Validation and Test Set

- **Training Set**: Subset of the full data used to find the parameters.
- **Validation Set**: Held-out subset of the full data used for model selection, i.e. finding the hyper-parameters.
- **Test Set**: Held-out subset of the data to estimate the test error of the best model.

Machine Learning Competitions e.g. on kaggle.com

1. Start of the competition: Participants obtain a data set, but not the test set.
2. Participants split the data set into training and validation sets as they want to fit the parameters and tune the hyper-parameters.
3. End of the competition: Organizers evaluate all submitted solutions on the test set.

Training, Validation and Test Set Q&A

- **Can I use the validation set to estimate the test error?**

Training, Validation and Test Set Q&A

- **Can I use the validation set to estimate the test error?**
  Yes.
- **Is the validation error of the winning model a good estimate of the test error?**
  No. The validation set estimate of the test error with the winning model tends to be too low, because the validation set was used for hyper-parameter tuning; the winning model may have had by chance a very low validation error.
- **Do I need a test set to find the best model?**

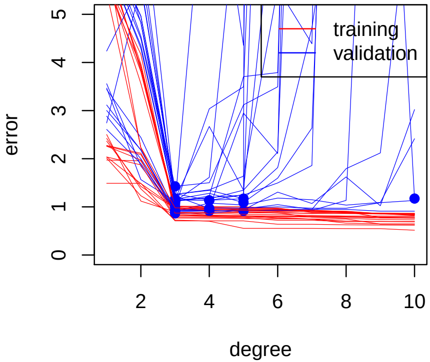
Training, Validation and Test Set Q&A

- **Can I use the validation set to estimate the test error?**
  Yes.
- **Is the validation error of the winning model a good estimate of the test error?**
  No. The validation set estimate of the test error with the winning model tends to be too low, because the validation set was used for hyper-parameter tuning; the winning model may have had by chance a very low validation error.
- **Do I need a test set to find the best model?**
  No. The test set is only used to estimate the test error. If you only care about finding the best model, but you do not care about its test error, you do not need a test set.
- **Should I fit my best model – the one with the winning hyper-parameters – on all available data on not just the training data?**

Training, Validation and Test Set Q&A

- **Can I use the validation set to estimate the test error?**
  Yes.
- **Is the validation error of the winning model a good estimate of the test error?**
  No. The validation set estimate of the test error with the winning model tends to be too low, because the validation set was used for hyper-parameter tuning; the winning model may have had by chance a very low validation error.
- **Do I need a test set to find the best model?**
  No. The test set is only used to estimate the test error. If you only care about finding the best model, but you do not care about its test error, you do not need a test set.
- **Should I fit my best model – the one with the winning hyper-parameters – on all available data on not just the training data?**
  Yes. To achieve best performance on unseen data, one should fit the winning model on all available data.

Drawback of Validation Set Approach



The validation estimate of the test error can be highly variable, depending on precisely which observations are included in the training set and which observations are included in the validation set $\Rightarrow$ high variance in model selection.

Quiz

Which of the following statements are correct?

- After finding in a model comparison the best performing model on the validation set, we compute the error on the validation set and the error on the test set.

Quiz

Which of the following statements are correct?

- After finding in a model comparison the best performing model on the validation set, we compute the error on the validation set and the error on the test set.
  - The test error is usually larger than the validation error.

Quiz

Which of the following statements are correct?

- After finding in a model comparison the best performing model on the validation set, we compute the error on the validation set and the error on the test set.
  - The test error is usually larger than the validation error.
  - Test error and validation error are roughly equal.

Quiz

Which of the following statements are correct?

- After finding in a model comparison the best performing model on the validation set, we compute the error on the validation set and the error on the test set.
  - The test error is usually larger than the validation error.
  - Test error and validation error are roughly equal.
  - The test error is usually smaller than the validation error.

Quiz

Which of the following statements are correct?

- After finding in a model comparison the best performing model on the validation set, we compute the error on the validation set and the error on the test set.
  - The test error is usually larger than the validation error.
  - Test error and validation error are roughly equal.
  - The test error is usually smaller than the validation error.
- The error on unseen data tends to be lower for a model trained on all available data compared to a model trained on a training set with 80% of all available data.

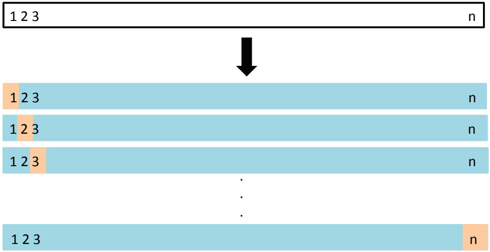
Quiz

Which of the following statements are correct?

- After finding in a model comparison the best performing model on the validation set, we compute the error on the validation set and the error on the test set.
    - The test error is usually larger than the validation error.
    - Test error and validation error are roughly equal.
    - The test error is usually smaller than the validation error.
- The error on unseen data tends to be lower for a model trained on all available data compared to a model trained on a training set with 80% of all available data.
- To have the most accurate models for model comparison it is acceptable to fit the models on all available data and compare them on the validation set consisting of 50% of the data.

Table of Contents

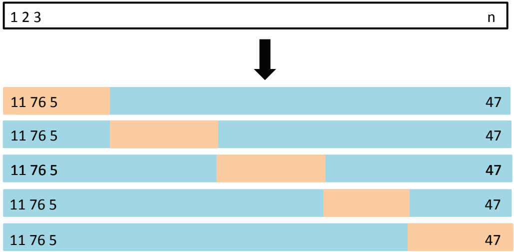
Leave-One-Out Cross-Validation (LOOCV)



$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{MSE}_i \qquad \mathrm{MSE_i} = (y_i - \hat{y}_i)^2$$

where $\hat{y}_i$ is the prediction obtained by fitting without (x_i, y_i).

K-Fold Cross-Validation

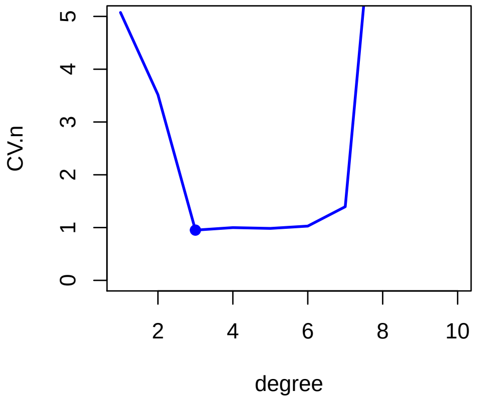


$$\mathrm{CV}_{(k)} = \sum_{k=1}^{K} \frac{n_k}{n} \mathrm{MSE}_k \qquad \mathrm{MSE}_\mathrm{k} = \frac{1}{n_k} \sum_{i \in C_k} (y_i - \hat{y}_i)^2$$

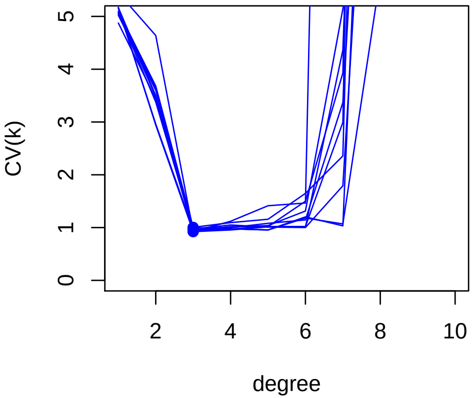
where $\hat{y}_i$ are predictions obtained by fitting without the data in part C_k.

Cross-Validation Applied to the Artificial Data

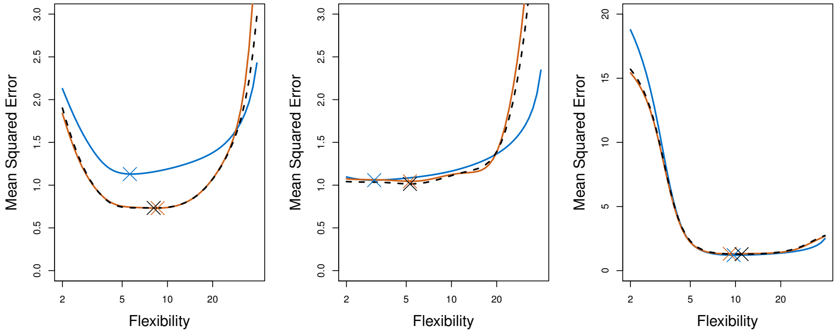
Leave-one-out Cross Validation



5-Fold Cross Validation



True versus Estimated Test Error



LOOCV (black dashed) and 10-fold CV (orange solid) find almost the same optimal flexibility as the true test error (blue). Crosses indicate the minima of the MSE curves.

Further Considerations

- The choice of the number of folds K is somewhat arbitrary. Typical choices are $K = 5$ or $K = 10$.

Further Considerations

- The choice of the number of folds K is somewhat arbitrary.
  Typical choices are $K = 5$ or $K = 10$.
- LOOCV has higher computational costs, since n fits are made instead of K
  (Except for least squares linear or polynomial regression.)

Further Considerations

- The choice of the number of folds K is somewhat arbitrary.
  Typical choices are $K = 5$ or $K = 10$.
- LOOCV has higher computational costs, since n fits are made instead of K
  (Except for least squares linear or polynomial regression.)
- To estimate the test error: fit the winner of model comparison to all data except the test set and evaluate it on the test set.

Further Considerations

- The choice of the number of folds K is somewhat arbitrary.
  Typical choices are $K = 5$ or $K = 10$.
- LOOCV has higher computational costs, since n fits are made instead of K
  (Except for least squares linear or polynomial regression.)
- To estimate the test error: fit the winner of model comparison to all data except the test set and evaluate it on the test set.
- To have the best model for predictions of future data: fit the model on all data you have.

Further Considerations

- The choice of the number of folds K is somewhat arbitrary.
  Typical choices are $K = 5$ or $K = 10$.
- LOOCV has higher computational costs, since n fits are made instead of K
  (Except for least squares linear or polynomial regression.)
- To estimate the test error: fit the winner of model comparison to all data except the test set and evaluate it on the test set.
- To have the best model for predictions of future data: fit the model on all data you have.
- If you do not care about an estimate of the test error, you run cross-validation on the full data, without first splitting off a test set.

Cross-Validation on Classification Problems

Instead of the log-likelihood one can also use the average misclassification rate on the held-out sample for cross-validation in classification problems.

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

Cross-Validation on Classification Problems

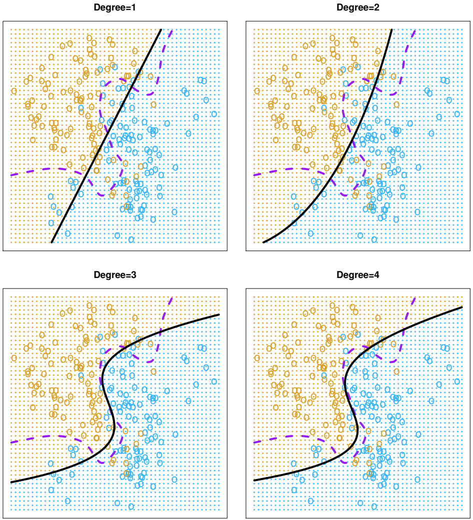
Instead of the log-likelihood one can also use the average misclassification rate on the held-out sample for cross-validation in classification problems.

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

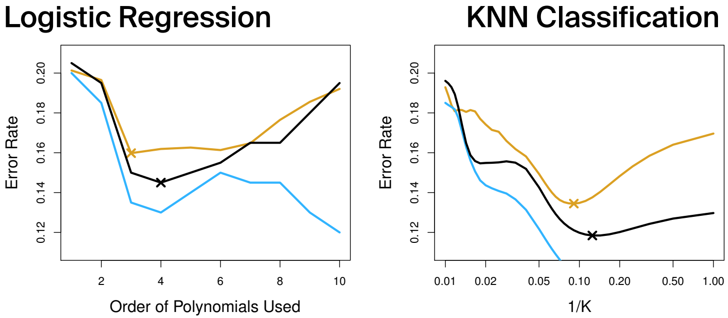
Optimal decision boundary (purple)

Estimates decision boundary (black)

for polynomial degrees 1-4.

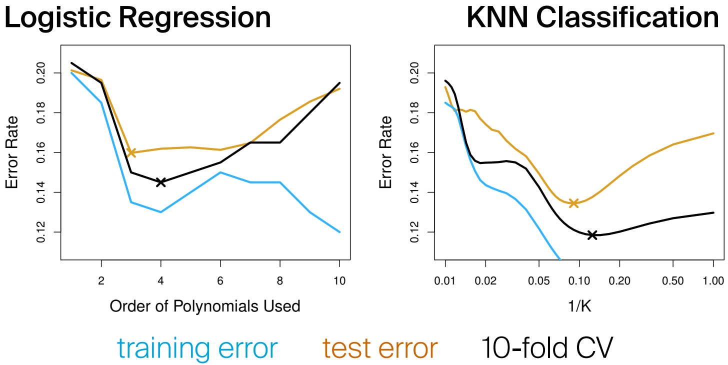


Cross-Validation on Classification Problems



training error test error 10-fold CV

Cross-Validation on Classification Problems



Logistic regression maximizes the likelihood of the parameters β_i given the data $\Rightarrow$ training likelihood is monotonically increasing with order of polynomials, but the training error (misclassification rate) is not necessarily decreasing.

Quiz

Which of the following statements are correct?

- Estimates of the test error with the validation set approach have lower variance than those with LOOCV.

Quiz

Which of the following statements are correct?

- Estimates of the test error with the validation set approach have lower variance than those with LOOCV.
- In a binary classification task we could use the AUC instead of the error rate to perform cross-validation.

Table of Contents

Tuning Models

Hyper-parameter tuning, i.e. finding the best model for the given data, **is an art.**

Common recipes:

- **Grid Search**: Perform cross-validation on a grid of hyper-parameter values. E.g. pick 10 different values of K and pick the best one with cross-validation.
- Use more sophisticated sampling methods for the hyper-parameters to be evaluated, see e.g. `https://www.automl.org/` or `https://github.com/baggepinnen/Hyperopt.jl`

Table of Contents

A Recipe for Supervised Learning

1. Collect (a lot of) data.

A Recipe for Supervised Learning

1. Collect (a lot of) data.
2. Look at the raw data; clean it if necessary.

A Recipe for Supervised Learning

1. Collect (a lot of) data.
2. Look at the raw data; clean it if necessary.
3. Select relevant features from the raw data,
   i.e. choose a suitable representation of the raw data.

A Recipe for Supervised Learning

1. Collect (a lot of) data.
2. Look at the raw data; clean it if necessary.
3. Select relevant features from the raw data,
   i.e. choose a suitable representation of the raw data.
4. Select a machine learning method.

A Recipe for Supervised Learning

1. Collect (a lot of) data.
2. Look at the raw data; clean it if necessary.
3. Select relevant features from the raw data,
   i.e. choose a suitable representation of the raw data.
4. Select a machine learning method.
5. Fit the data and tune hyperparameters, e.g. with cross-validation.

A Recipe for Supervised Learning

1. Collect (a lot of) data.
2. Look at the raw data; clean it if necessary.
3. Select relevant features from the raw data,
   i.e. choose a suitable representation of the raw data.
4. Select a machine learning method.
5. Fit the data and tune hyperparameters, e.g. with cross-validation.

A Recipe for Supervised Learning

1. Collect (a lot of) data.
2. Look at the raw data; clean it if necessary.
3. Select relevant features from the raw data,
   i.e. choose a suitable representation of the raw data.
4. Select a machine learning method.
5. Fit the data and tune hyperparameters, e.g. with cross-validation.
6.
   - training loss: high, test loss: high (underfitting?): select a more flexible method.

A Recipe for Supervised Learning

1. Collect (a lot of) data.
2. Look at the raw data; clean it if necessary.
3. Select relevant features from the raw data,
   i.e. choose a suitable representation of the raw data.
4. Select a machine learning method.
5. Fit the data and tune hyperparameters, e.g. with cross-validation.
6.
   - training loss: high, test loss: high (underfitting?): select a more flexible method.
   - training loss: low, test loss: high (overfitting?): select a less flexible method.

A Recipe for Supervised Learning

1. Collect (a lot of) data.
2. Look at the raw data; clean it if necessary.
3. Select relevant features from the raw data,
   i.e. choose a suitable representation of the raw data.
4. Select a machine learning method.
5. Fit the data and tune hyperparameters, e.g. with cross-validation.
6.
   - training loss: high, test loss: high (underfitting?): select a more flexible method.
   - training loss: low, test loss: high (overfitting?): select a less flexible method.
7. Repeat 4-6 until the lowest test loss is found.

A Recipe for Supervised Learning

1. Collect (a lot of) data.
2. Look at the raw data; clean it if necessary.
3. Select relevant features from the raw data,
   i.e. choose a suitable representation of the raw data.
4. Select a machine learning method.
5. Fit the data and tune hyperparameters, e.g. with cross-validation.
6.
   - training loss: high, test loss: high (underfitting?): select a more flexible method.
   - training loss: low, test loss: high (overfitting?): select a less flexible method.
7. Repeat 4-6 until the lowest test loss is found.
8. If unhappy with the lowest test loss, repeat 2-7 or collect more data.

A Recipe for Supervised Learning

1. Collect (a lot of) data.
2. Look at the raw data; clean it if necessary.
3. Select relevant features from the raw data,
   i.e. choose a suitable representation of the raw data.
4. Select a machine learning method.
5. Fit the data and tune hyperparameters, e.g. with cross-validation.
6. - training loss: high, test loss: high (underfitting?): select a more flexible method.
   - training loss: low, test loss: high (overfitting?): select a less flexible method.
7. Repeat 4-6 until the lowest test loss is found.
8. If unhappy with the lowest test loss, repeat 2-7 or collect more data.
9. Fit the best model on all available data for best performance on unseen data.