Overfitting and Model Selection

Machine Learning Primer Course

Georgios C. Anagnostopolous and Spencer G. Lyon

October 2, 2020

Purpose

- Explain what model over-fitting is and when it occurs
- Explain the need for model selection and how it is achieved via validation procedure
- Discuss 3 common types of validation procedures

Table of contents

1. Model Overfitting

2. Model Selection

3. Glossary

Model Overfitting

What is Overfitting?

- Low training MSE $\implies$ trained model fits the training data well
  - The fitted residuals are all small $\implies$ the fitted response hyper-surface passes very close to targets
  - Training MSE is zero $\implies$ hyper-surface passes through all training targets

---

[1] Often called "out-of-sample data" or "hold-out data"

[2] Hold-out MSE is the MSE computed on a hold-out set using the fitted model, i.e. the model using optimal weights

What is Overfitting?

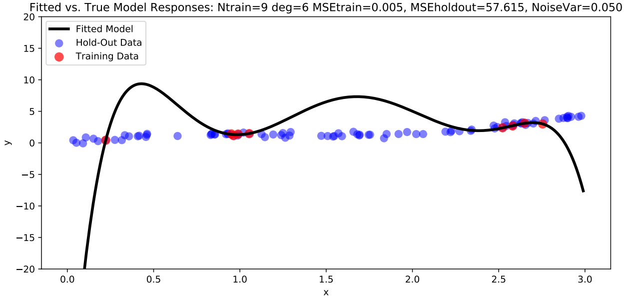
- Low training MSE $\implies$ trained model fits the training data well
  - The fitted residuals are all small $\implies$ the fitted response hyper-surface passes very close to targets
  - Training MSE is zero $\implies$ hyper-surface passes through all training targets
- Generalization performance of the fitted model may be unacceptably bad
  - Predictions for samples not used in training[1] are poor
  - Hold-out MSE[2] is orders of magnitude larger than training MSE

[1] Often called "out-of-sample data" or "hold-out data"

[2] Hold-out MSE is the MSE computed on a hold-out set using the fitted model, i.e. the model using optimal weights

Overfitting Example

- Polynomial regression with 9 training samples and 6th degree polynomial



- Observations
  - Perfect fit of training data => small training MSE
  - Poor fit of out-of-sample data => large hold-out MSE

Overfitting in Linear Regression

- Assume N training samples and P weights
- When $P \geq N$ the model is over-parameterized[3]
  - The fitted model will exactly fit training data $\implies \text{MSE} = 0$
  - When $P = N$ there is a unique $\mathbf{w}^*$
  - When $P > N$ there are multiple solutions for $\mathbf{w}^*$
    - Matrix $\mathbf{R}$ is not invertible; a solution can be found via pseudo-inverse of $\mathbf{X}$
- As N becomes larger than P, overfitting gradually subsides

[3] Even when $N > P$, but "close" we will still use this terminology

Overfitting in Linear Regression

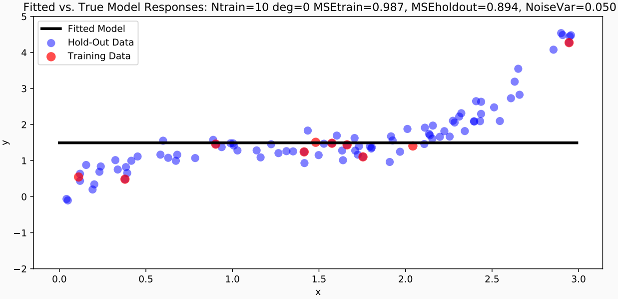
- Another way to look at it ...
  - When a model is over-parameterized relative to training set, it is "flexible" enough to yield a perfect or very good fit on training data
- Countering overfitting
  - Use more training data
  - If not possible, use fewer features to reduce P
  - If not feasible, use a regularization approach[4]

[4] To be discussed in an upcoming lecture

The flip side

Model underfitting:

- Both training MSE and hold-out MSE are relatively "large"
- If one used a more flexible model, both MSEs would fall
- In this scenario the model is under-parameterized



Model Selection

Model Selection

- Select between a set of models with varying flexibility[5]
- Goal: Choose a model that has the best generalization properties

[5] Flexibility could be increased in a variety of ways including using a different functional form for the model, adding additional features, transforming existing features, etc.

Model Selection

- Select between a set of models with varying flexibility[5]
- Goal: Choose a model that has the best generalization properties
- Example: polynomial regression
  - Varying number of features, controlled by degree D of polynomial
  - The number D indexes a family of regression models and controls flexibility
  - Optimal D cannot be determined based solely on training set (beware of overfitting!)
  - Parameters like D are called hyper parameters

[5] Flexibility could be increased in a variety of ways including using a different functional form for the model, adding additional features, transforming existing features, etc.

Model Selection continued ...

Main point

- Model selection cannot be solely based on the training set.
- Training MSE informs us about model fit, but not necessarily about the model's generalization performance (not trustworthy)
- A (more) honest loss estimate needs to be employed
  - We used a hold-out set for this purpose (more trustworthy)

Validation Procedure

- An approach to model selection: out of a pool of candidate models, guess which one may exhibit the best generalization (let's decide to call it the champion model).

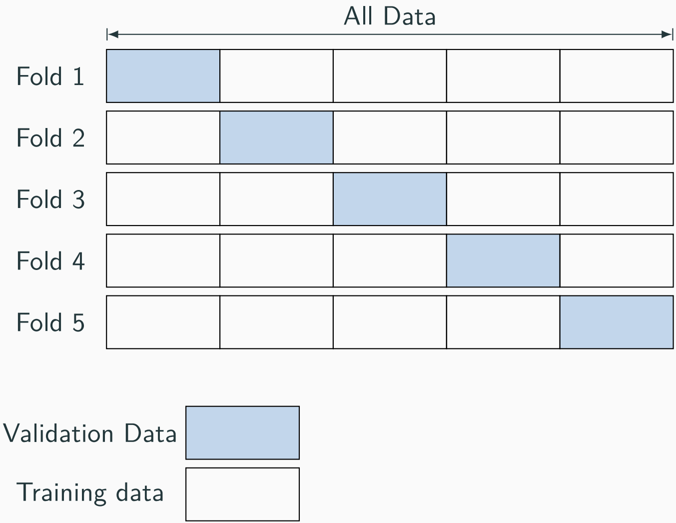
Validation Procedures

- There are many, we'll only cover a few
- Hold out method
  - Partition available data into training (typically, larger set) and (typically, smaller set) hold-out set, called the validation set.
  - Fit models on training set
  - Select model with best average loss on hold-out set
  - Considerations
    - Large training set $\Longrightarrow$ better fitting models
    - The larger the validation set, the better quality of the generalization estimate (e.g. $\mathrm{MSE}_{\mathrm{val}}$)
    - Assumes there are plenty of data for both sets

K-fold Cross Validation

- Partition available data into K equally-sized subsets (folds).
- For $k = 1$ to K
  - Pick a fold to play the role of the validation set
  - Use the remaining $K - 1$ folds for training the models
  - Compute the average loss on the validation set/fold
- Use the sample average of the $\left\{\mathrm{MSE}_{\mathrm{val}}^{k}\right\}_{k=1}^{K}$ to guess the best generalizing model (champion)

Visualizing K-fold CV



K-fold CV

- Considerations for K-fold CV
  - For fixed number of available data N
    - large $K \implies$ large training sets and small validation sets $\implies$ worse quality of generalization performance estimate
  - Used when data are deemed "not enough" to employ hold-out method
  - Often, people use K=10 (although completely arbitrary)
- Leave ont out cross validation (LOOCV)
  - K-fold CV to the extreme: $K = N$
  - Considerations
    - Used when available data are "too few"
    - Training sets almost identical from fold to fold $\implies$ trained models typically differ by very little from fold to fold

LOOCV with Linear Regression

An amazing thing...

- The LOOCV estimate for linear regression has a closed form based on the diagonal elements of the hat matrix $\mathbf{H}$.

$$\text{MSE}_{\text{loocv}} = \frac{1}{N}\sum_{n=1}^{N}\frac{\hat{e}_n^2}{(1-h_{n,n})^2} = \frac{1}{N}\hat{\mathbf{e}}^{*T}\left(\mathbf{I}_N - \mathbf{H}_{\text{diag}}\right)^{-2}\hat{\mathbf{e}}^*$$

$$\mathbf{H} \triangleq \mathbf{X}\mathbf{X}^\dagger$$

$$\mathbf{H}_{\text{diag}} \triangleq \text{diagonal elements of } \mathbf{H} = \begin{bmatrix} h_{1,1} & 0 & \cdots \\ 0 & h_{2,2} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

- In the literature it is called the Predicted Residual Error Sum of Squares (PRESS) statistics
- In essence it is a weighted form of the training MSE

Comments about Model Selection

- Hold-out method more reliable than K-fold CV or LOOCV
  - It will provide most honest estimate of generalization performance
  - However, hold-out data is not used for training...
  - It is most commonly used when there is ample data
- In general, the best validation procedure depends on
  1. The context of the ML task
  2. The size of available data set
  3. Efficiency/feasibility concerns[6]
  4. The choice of hyper-parameters (such as the number of folds K in K-fold CV)
- This is a difficult problem that is being actively researched

[6] If it takes many days to run one fold of the CV procedure, perhaps you can't afford to have many folds

Practical Advice

So...which to choose?

- Hard to tell and often largely arbitrary[7]
- In practice, a trial and error approach is used
- Main considerations
  - Choose the method that allows for enough training samples to get a good-fitting model. Goal: make training MSE similar to validation MSE
  - Time complexity of validation procedure: LOOCV (i.e. N-fold cross-validation) is, in general, more computational expensive than K-fold cross-validation for $K < N$[8]

---

[7] Though, in a research setting researchers are typically required to provide a reasonable defense of their choice

[8] Exception: for linear regression, the opposite holds thanks to the closed-form of the validation MSE

Types of Candidate Pools

- Recall that validation procedures begin with defining a set of candidate models
- We call this the candidate pool
- Common types of candidate pools are:
  - Models of the same structure that consider different sets of features (here validation helps in feature selection)

Types of Candidate Pools

- Recall that validation procedures begin with defining a set of candidate models
- We call this the candidate pool
- Common types of candidate pools are:
  - Models of the same structure that consider different sets of features (here validation helps in feature selection)
  - Family of models indexed by one or more hyper parameters (e.g. structure of neural network – here validation addresses model regularization)

Types of Candidate Pools

- Recall that validation procedures begin with defining a set of candidate models
- We call this the candidate pool
- Common types of candidate pools are:
  - Models of the same structure that consider different sets of features (here validation helps in feature selection)
  - Family of models indexed by one or more hyper parameters (e.g. structure of neural network – here validation addresses model regularization)
  - A potpourri collection of models not belonging to a common family – called a bag of models, for lack of a better term

The Test Set

Consider...

- You have selected a pool of models
- Applied a validation procedure to select the model and hyper parameters
- Now you want to get a sense of ”true” generalization performance
- How could you do this?

The Test Set

Consider...

- You have selected a pool of models
- Applied a validation procedure to select the model and hyper parameters
- Now you want to get a sense of ”true” generalization performance
- How could you do this?
- Answer: Use a third type of set: test set
- Test set not used for training (model fitting) or validation (model selection)

Glossary

Comments on Nomenclature

- Unfortunately, there is no consensus in precisely defining what a validation, hold-out and test set is
- Hence, when these terms come up in a validation procedure description (i.e. in a paper or a software implementation), one has to figure out what is meant by these terms

Our Conventions (1/3)

Hyper parameter
A parameter that indexes a family of models. Its optimal value cannot be meaningfully determined from the training set, i.e. by minimizing the training loss, but is determined through a validation procedure

Our Conventions (1/3)

Hyper parameter
A parameter that indexes a family of models. Its optimal value cannot be meaningfully determined from the training set, i.e. by minimizing the training loss, but is determined through a validation procedure

Example: Consider the context of polynomial regression.

- Consider a pool (family) of polynomials of varying degree D in $\{0, \ldots, D_{\text{max}}\}$
- We cannot determine a best D^* (in terms of generalization) by minimizing the training MSE
- The training MSE is always minimized when $D^* = D_{\text{max}}$
- However, we are very likely to overfit and have poor generalization
- In this example, D is a hyper parameter, whose best value needs to be determined through a validation procedure

Our Conventions (2/3)

Validation procedure
A model selection procedure that identifies the champion model among a pool of candidate models under consideration that most likely will exhibit the best generalization. Often, these models will be indexed by one or more hyper-parameters; hence, the task becomes identifying the best hyper-parameter value(s) to use for training a well-generalizing model.

Hold-out set
A generic term used for any set of samples not used for training model(s)

Validation set
A subset of hold-out data, used to assess the generalization potential of models[9]

[9] All 3 validation procedures we discussed use sets in such capacity

Our Conventions (3/3)

Validation MSE
An MSE estimate computed by a validation procedure.

- For the hold-out method, it is the MSE as computed on the one and only validation set.
- For K-fold cross-validation (and, hence, for LOOCV as well, since it is a special case), it is the sample average of all MSEs computed on the different folds (which act as validation sets).

Test set
A set of samples that is neither used for training nor for validation purposes. Such a set is used to get an honest estimate of how good a model (e.g. a champion model) generalizes. In the wild, test sets and validation sets are often conflated terms.