Cross validation

JrPhy

Introduction

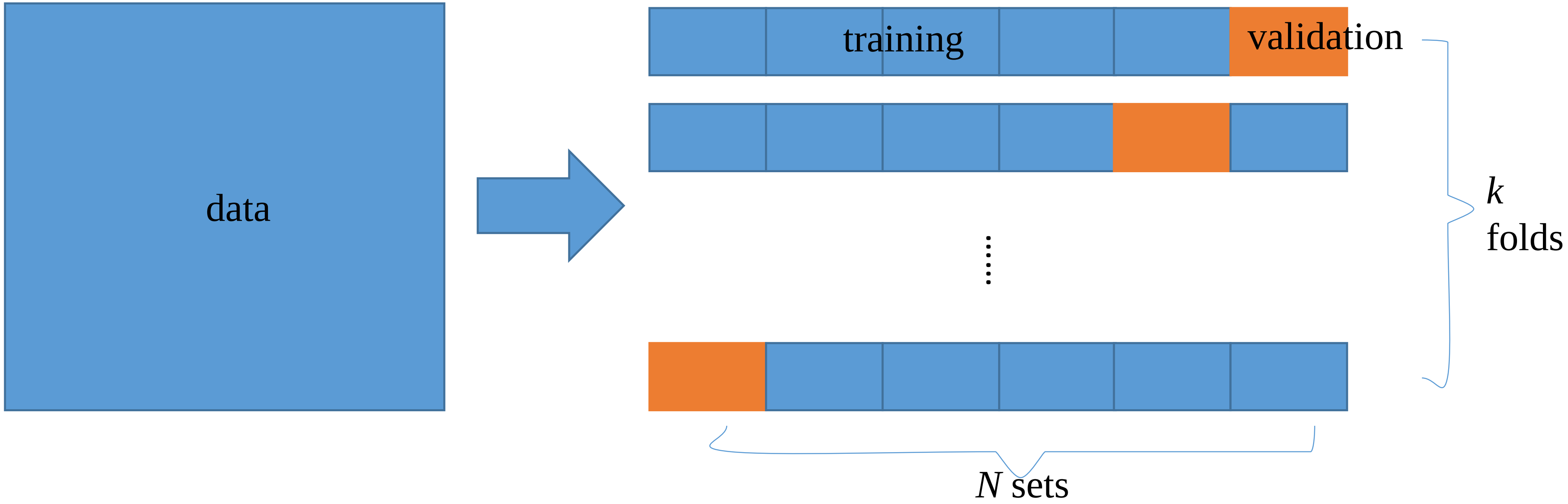
- So far I’ve introduced so many algorithm for machine learning, and each algorithm can generate so many models. So here I want to discuss how to evaluate and choose the best model.
- The step of machine learning is
- 1. collect data
- 2. select an algorithm and train a model
- 3. predict in practical
- So we want to get a accuracy model to predict the correct data.

Cross validation

- There are 2 error we focus on them, one is E_{in}, it bases on the model which is trained by the data has been collected. Suppose $f\,(x_{collect})$ is the model trained by the data, $y_{collect}$ is the answer, then
- $E_{in} = \|\, y_{collect} - f\,(x_{collect})\, \|$ or in other forms.
- The other is E_{out}, it bases on the model has been trained for predicting. Suppose $f\,(x_{predict})$ is the model trained by the data, $y_{predict}$ is the answer, then
- $E_{out} = \|\, y_{predict} - f\,(x_{predict})\, \|$ or in other forms.

Cross validation(tuning parameters)

- If the all data are used for training, then we can just know the E_{out} when you predict the result, it's very dangerous for practical use.
- So we separate data into k folds, and each fold is separated into N sets, one is validation set, and the others are training set.



Cross validation(tuning parameters)

- Use each training set to get a model, $f_{(N-1)}^{(i)}$, the symbol means f is trained by the $(N-1)$ sets, then use the validation set to get $E_{val}(f_{(N-1)}^{(i)})$, so the criteria of choosing model is minimum E_{val}. So use the hypothesis of minimum E_{val} , and all data to train a new model f, then it may be the best model in such the problem, and you can use the result to pursue other people that your model is robust.
- Each fold of data comes from the original data, so that it can be seem as in the same distribution.

How many folds?

- Denotes the error of cross-validation is $C[f_{(N-1)}^{(i)}]$, its average is C_{cv}. Suppose $C[f_{(N-1)}^{(i)}]$ is the estimator is $\mathrm{E}_\mathrm{X}(C[f_{(N)}])$, then so as C_{cv}, so
  - MSE(C_{cv}) = $\mathrm{E}_\mathrm{X}\,[(C_{cv} - E_X\,(C\,[f_{(N)}]))^2]$, = $\mathrm{Var}_\mathrm{X}\,(C_{cv})$ + bias $(C_{cv})^2$

- bias (C_{cv}) = $\mathrm{E}_\mathrm{X}\,(C_{cv}) - \mathrm{E}_\mathrm{X}(C[f_{(N)}])$

$$= E\left(\sum_{i=1}^{k} \frac{1}{k} C[\,f_{(N-1)}{}^{(i)}\,]\right) - E(C[\,f_N\,])$$

$$= \frac{1}{k}\sum_{i=1}^{k} E\Big(C[\,f_{(N-1)}{}^{(i)}\,]\Big) - E(C[\,f_N\,])$$

$$= E\Big(C[\,f_{(N-1)}{}^{(i=s)}\,]\Big) - E(C[\,f_N\,])$$

- = bias$(C[f_{(N-1)}^{(s)}])$ $\forall s$

How many folds?

$$\mathrm{Var}_{\mathrm{X}}(C_{cv}) = \mathrm{Var}\left(\sum_{i=1}^{k}\frac{1}{k}C[\,{f_{(N-1)}}^{(i)}]\right) = \frac{1}{k^{2}}\,\mathrm{Var}\left(\sum_{i=1}^{k}C[\,{f_{(N-1)}}^{(i)}]\right)$$

$$= \frac{1}{k^{2}}\,\mathrm{Var}\left(\sum_{i=1}^{k}C[\,{f_{(N-1)}}^{(i)}]\right) + 2\sum_{j>i}^{k}\sum_{i=1}^{k}\mathrm{Cov}_{\mathrm{X}}\left(C[\,{f_{(N-1)}}^{(i)}], C[\,{f_{(N-1)}}^{(i)}]\right)$$

$$= \frac{1}{k^{2}}\,\mathrm{Var}\left(C[\,{f_{(N-1)}}^{(s)}]\right) + \frac{2}{k^{2}}\sum_{j>i}^{k}\sum_{i=1}^{k}\mathrm{Cov}\left(C[\,{f_{(N-1)}}^{(i)}], C[\,{f_{(N-1)}}^{(i)}]\right) \ \forall s$$

How many folds?

More data
set to reduce

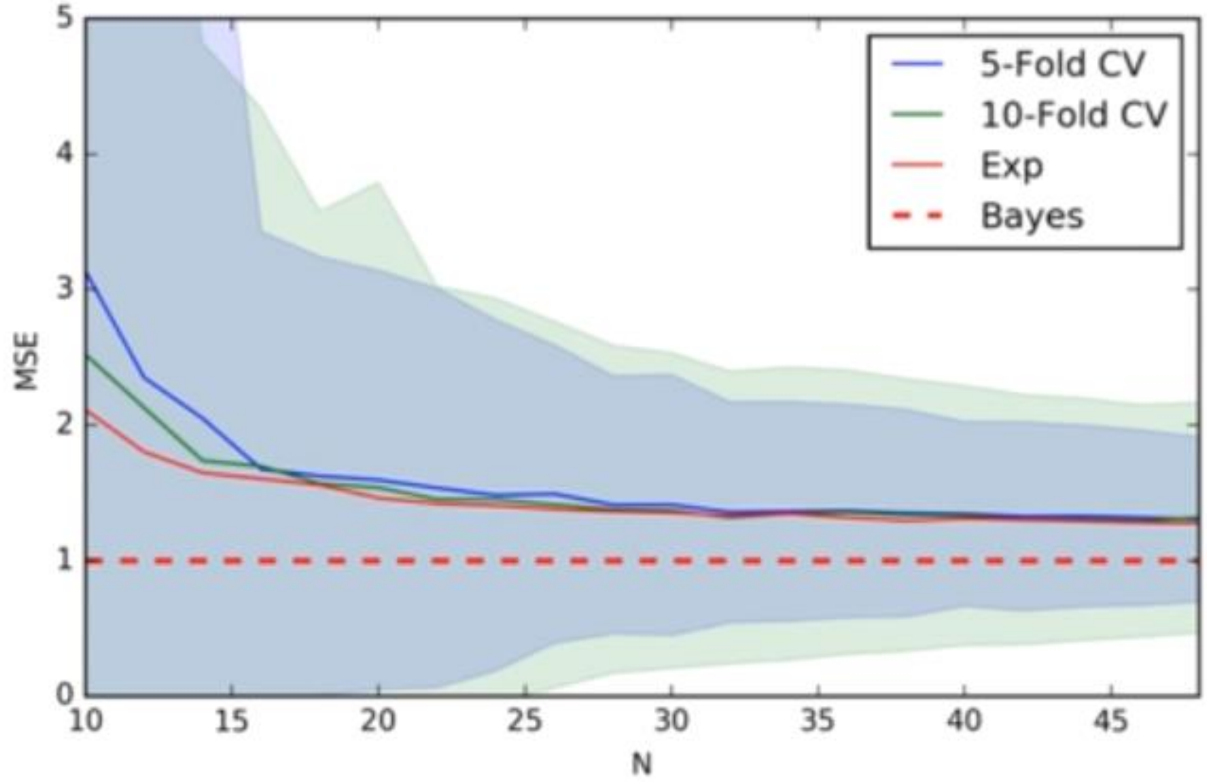
- By previous derivation, we know
- bias $(C_{cv}) = \boxed{\text{bias}(C[f_{(N-1)}{}^{(s)}])}$, and

Less data set
to reduce

$$\text{Var}_X(C_{cv}) = \boxed{\frac{1}{k}\text{Var}\left(C[f_{(N-1)}{}^{(s)}]\right)} + \boxed{\frac{2}{k^2}\sum_{j>i}^{k}\sum_{i=1}^{k}\text{Cov}\left(C[f_{(N-1)}{}^{(i)}], C[f_{(N-1)}{}^{(i)}]\right)} \quad \forall s$$

- We can reduce the bias and Var by choosing correct model complexity to avoid the under- and over-fitting, or less data in the validation fold that is, more folds in a set are used for training.
- The other is the correlation term, if the folds are more, it’s easier to get the same data, so the correlation increases as the number of folds increases.
- So in practical, k = 5 or 10 frequently.

MSE versus N



Shan-Hung Wu (CS, NTHU) CV & Ensembling Machine Learning 14 / 34

https://www.youtube.com/watch?v=G-SbnfqRw14&list=PLlPcwHqLqJDnGHwQCumd-gKDMt3NALHeY&index=2

Leave-One-Out CV(LOOCV)

- In the extreme case, how about there is just a datum in the validation fold? It means the collected data is very small, then the by the previous slide, the bias and Var are high, so use more data for training, the largest number is $k = N$, that is, only one datum for validation.
- This case is just for the **small** data set.