Training error, Test error

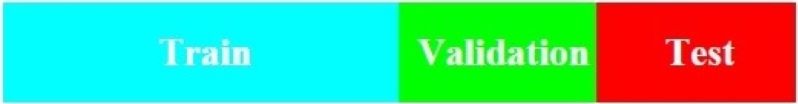
Instructor: Farzaneh Mirzazadeh

Department of Computer Science, UCSC, Winter 2017

Announcements

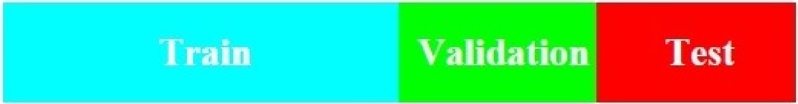
- Delivery of assignment
- Late policy details
- Last submission is counted

How to set the regularization hyper parameter β?



- Split data: Train, Test (say 80%, 20%)
- Use a portion of training set, as a validation set.
- Do not train on validation set. Only evaluate a number of different values of regularization parameter, and find the smallest that has a better performance on validation set.
- Which values to check? A Grid with logarithmic scale could be a good idea.
- The parameter selecting should not have any information about test data to be fair.

How to set the regularization hyper parameter β?



- Split data: Train, Test (say 80%, 20%)
- Use a portion of training set, as a validation set.
- Do not train on validation set. Only evaluate a number of different values of regularization parameter, and find the smallest that has a better performance on validation set.
- Which values to check? A Grid with logarithmic scale could be a good idea.
- The parameter selecting should not have any information about test data to be fair.

Is it possible to use both L_1 and L_2 regularization

- Yes, e.g. elastic net combines both regularizers with sum of squares error.
- But, L_1 blocks representer theorem. will not have kernels.
- There are (evolved) ways of having both kernels and sparsity. (Out of scope of this course.)

Is it possible to use both L_1 and L_2 regularization

- Yes, e.g. elastic net combines both regularizers with sum of squares error.
- But, L_1 blocks representer theorem. will not have kernels.
- There are (evolved) ways of having both kernels and sparsity. (Out of scope of this course.)

Is it possible to use both L_1 and L_2 regularization

- Yes, e.g. elastic net combines both regularizers with sum of squares error.
- But, L_1 blocks representer theorem. will not have kernels.
- There are (evolved) ways of having both kernels and sparsity. (Out of scope of this course.)

Is it possible to use both L_1 and L_2 regularization

- Yes, e.g. elastic net combines both regularizers with sum of squares error.
- But, L_1 blocks representer theorem. will not have kernels.
- There are (evolved) ways of having both kernels and sparsity. (Out of scope of this course.)

Generalization

- Machine learning has a difference with optimization.
- It is important that the model performs well on new unseen data.
- Generalization
- Need assumptions
- IID assumption (Independent and identically distributed)
  - Examples in each data set are independent from each other.
  - Training and test set are drawn from the same probability distribution as each other.

Generalization

- Machine learning has a difference with optimization.
- It is important that the model performs well on new unseen data.
- Generalization
- Need assumptions
- IID assumption (Independent and identically distributed)
  - Examples in each data set are independent from each other.
  - Training and test set are drawn from the same probability distribution as each other.

Generalization

- Machine learning has a difference with optimization.
- It is important that the model performs well on new unseen data.
- Generalization
- Need assumptions
- IID assumption (Independent and identically distributed)
  - Examples in each data set are independent from each other.
  - Training and test set are drawn from the same probability distribution as each other.

Generalization

- Machine learning has a difference with optimization.
- It is important that the model performs well on new unseen data.
- Generalization
- Need assumptions
- IID assumption (Independent and identically distributed)
  - Examples in each data set are independent from each other.
  - Training and test set are drawn from the same probability distribution as each other.

Generalization

- Machine learning has a difference with optimization.
- It is important that the model performs well on new unseen data.
- Generalization
- Need assumptions
- IID assumption (Independent and identically distributed)
  - Examples in each data set are independent from each other.
  - Training and test set are drawn from the same probability distribution as each other.

Generalization

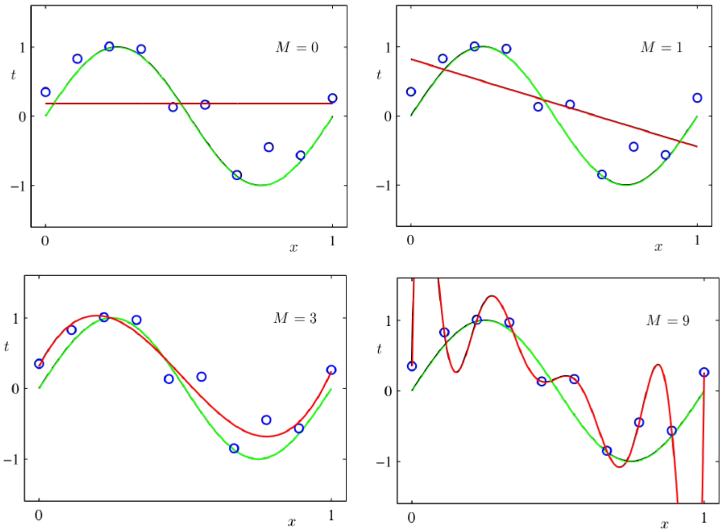
- Machine learning has a difference with optimization.
- It is important that the model performs well on new unseen data.
- Generalization
- Need assumptions
- IID assumption (Independent and identically distributed)
  - Examples in each data set are independent from each other.
  - Training and test set are drawn from the same probability distribution as each other.

Generalization

- Machine learning has a difference with optimization.
- It is important that the model performs well on new unseen data.
- Generalization
- Need assumptions
- IID assumption (Independent and identically distributed)
  - Examples in each data set are independent from each other.
  - Training and test set are drawn from the same probability distribution as each other.

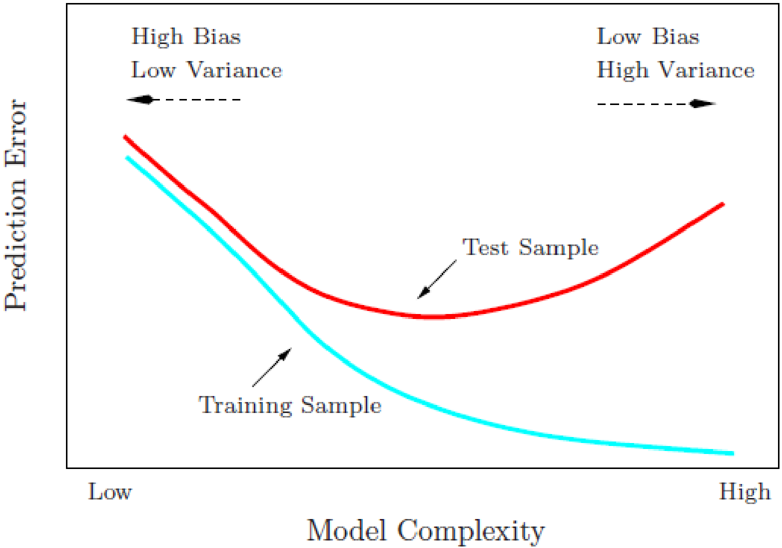
Complexity of hypothesis:

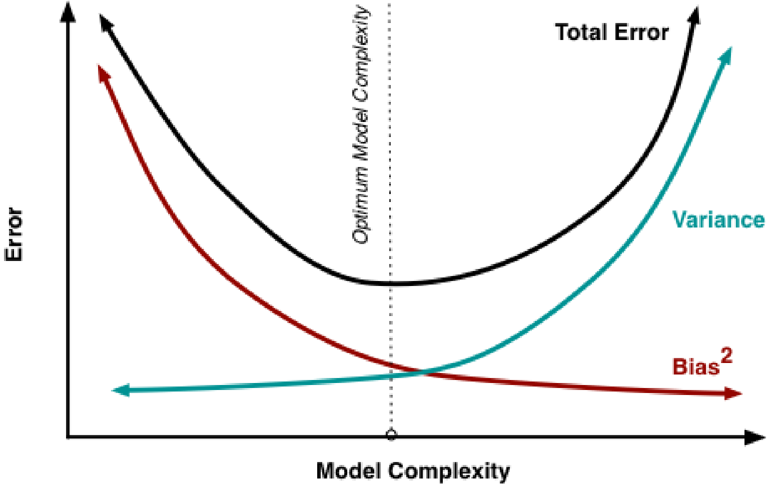
Bishop fig 1.4



17

Train vs Test Error




Error
Optimum Model Complexity
Total Error
Variance
Bias^2
Model Complexity