

Big Data Analysis

Lecture 8

2019/12/9

Logistic Regression

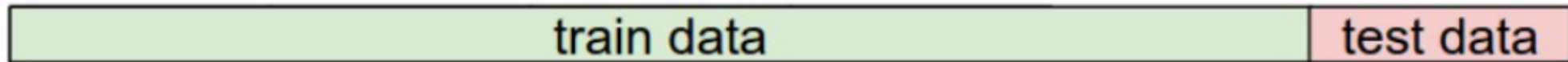
- Comparison with linear regression
- Generative models vs discriminative models
- Materials are borrowed and modified from 2 resources
 - Machine Learning from Hung-yi Lee, http://speech.ee.ntu.edu.tw/~tlkagk/courses_ML17_2.html
 - CS 109 Data Science from Harvard, <http://cs109.github.io/2015/>

Generative vs. Discriminative Models

- Training classifiers involve estimating $f: X \rightarrow Y$, or $P(Y | X)$
- Generative classifiers
 - Assume some functional form for $P(Y)$, $P(X | Y)$
 - Estimate parameters of $P(X | Y)$, $P(Y)$ directly from training data
 - Use Bayes rule to calculate $P(Y | X)$
- Discriminative Classifiers
 - Assume some functional form for $P(Y | X)$
 - Estimate parameters of $P(Y | X)$ directly from training data

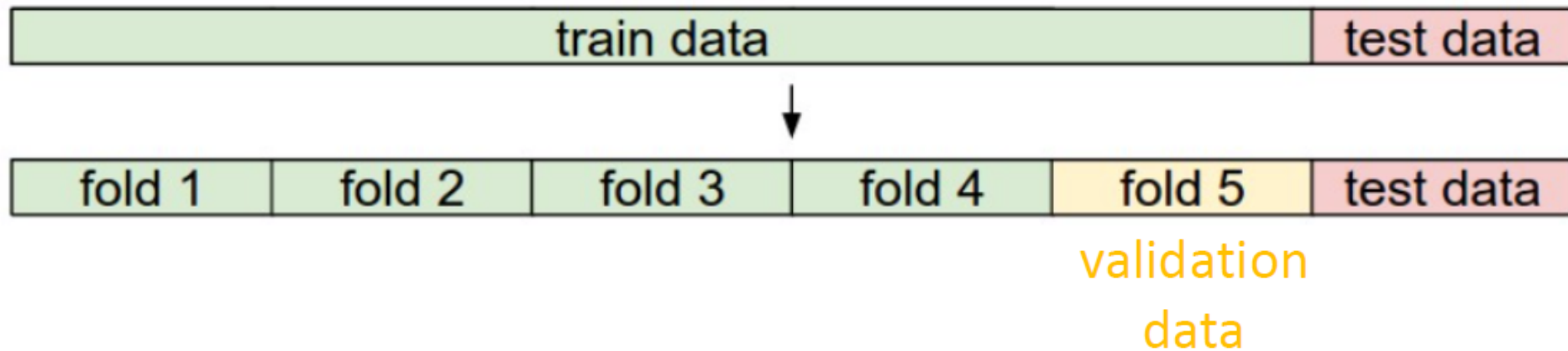
Validation and Resampling

- Train on training data, test on test data
- Pick the k with the lowest test error



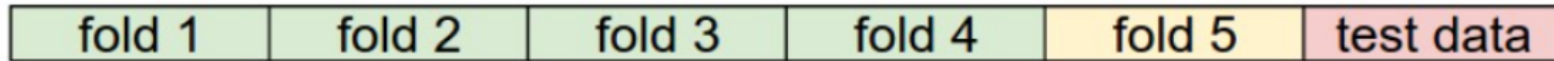
- **Training data**: train classifier
- **Test data**: measure performance

Cross Validation



- **Training data:** train classifier
- **Validation data:** estimate (hyper) parameters (k)
- **Test data:** measure performance





Cross Validation



1. Iterate over choice of validation fold
2. For all parameter values:
 - a. Train with training data
 - b. Validate with validation data
3. Average the parameters with best performance on validation data

The test data is **NOT** used to determine the parameters!

Cross Validation

- Take best parameters 
- Train on training data and validation data  
together
- Test performance on test data 

Evaluate on the test set only a **single time**, at the very end!

Crossing Validation and Resampling

- Pros
 - No parametric or theoretic assumptions
 - Given enough data, highly accurate
 - Conceptually simple
- Cons
 - Computationally intensive
 - Must choose the fold size
 - Potential conservative bias

<http://scott.fortmann-roe.com/docs/MeasuringError.html>

FX Market Making Example

- Retail Trading 24x5
 - Begins with Sydney session at 9am Monday (Sunday 10pm GMT)
 - Ends with NYC session at 5pm Friday (Friday 10pm GMT)
- Three sites: London, NYC, Singapore
- G10 Currencies:
 - USD, EUR, GBP, JPY, AUD, NZD, CAD, CHF, NOK, SEK