CISC 372
Advanced Data Analytics
L4 – Model Selection

https://l1nna.com/372

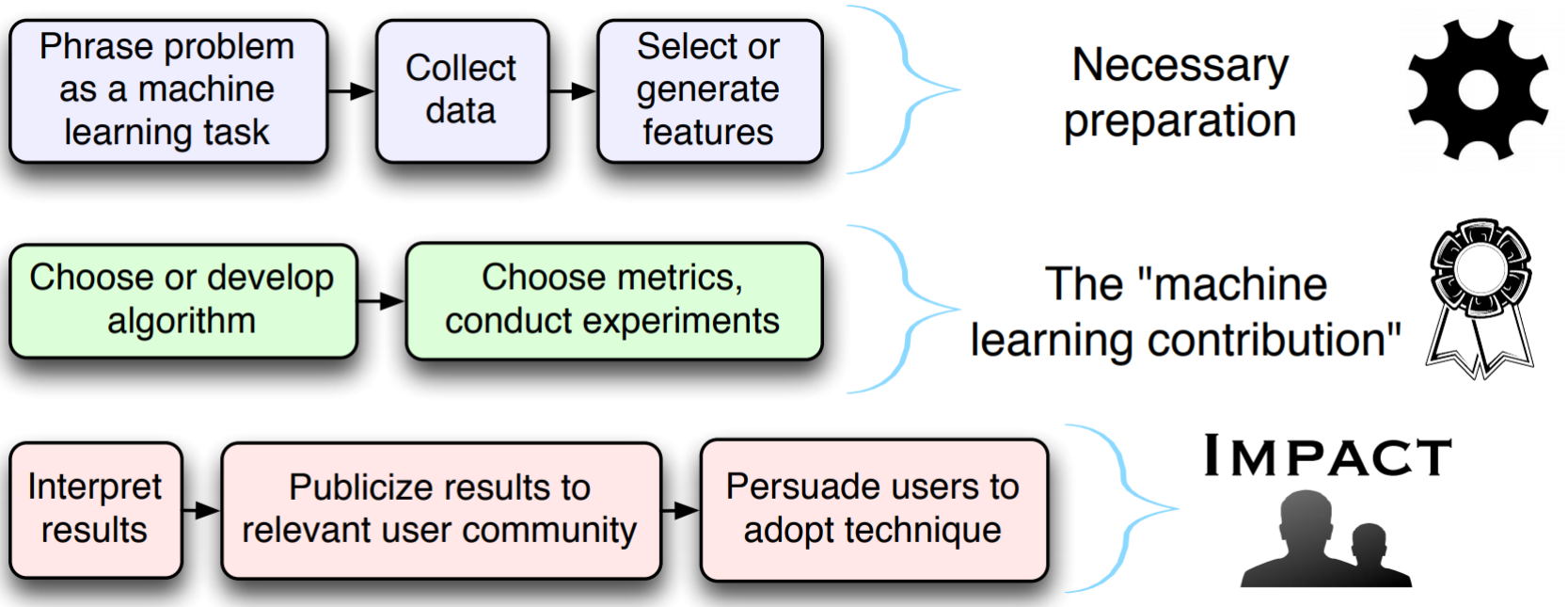
Timeline (revised.)

- Working with data records (Jan.)
  - Intro, review of linear methods, ethical/security/privacy
  - Model tuning and experimental protocol, data preprocessing & visualization, ensemble method
  - decision tree + random forest, xgboosting, CNN, tensor-based transformation,
  - instance-based learning, Bayesian method, density-based clustering

- Working with sequential/set-based data (Feb.)
  - use cases & data preprocessing, representation, visualization
  - Time series statistical learning, Association rule mining, Aprior Algorithm & FP-Tree
  - Sequential Data Mining, NLP, RNN + attention mechanism,
  - graphical model: topic modeling, word/paragraph embedding, BERT

- Working with graph data (Mar.)
  - From social network to heterogeneous information network (HIN), preprocessing and visualization, use cases
  - Network-based statistical modeling, HIST & Page Rank, Recommendation System
  - Community detection, Graph embedding, DeepWalk, Metapath2vec
  - LINE, GraphSAGE, presentation

Data Science Project



https://www.wkiri.com/research/papers/wagstaff-MLmatters-12.pdf

Last week

- Data Science/Data Analytics
  - A new approach to model complex system
  - Empirical science/Theoretical science/Computational Science/Data Science
- Review of Optimization-based Methods
  - Linear regression
  - Logistic regression
  - K-mean & K-medoids
- Ethical issues, privacy, and security concerns
  - CIA model for threat identification
  - AI/DS – Increased degree of uncertainty

Today

- Key terminologies
  - Supervised vs Unsupervised Learning
  - Classification vs Prediction
  - Generative model vs Discriminative model
- Error Estimation
  - Performance Metric
  - Bias-Variance trade-off
  - Experimental protocol
    - Simple hold-out
    - Cross-validation
    - Boostraping
    - Hold-out with validation/development set

Supervised vs. Unsupervised

- Supervised learning (classification)
  - Supervision: The training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations
  - New data is classified by the model constructed from the training set
- Unsupervised learning (clustering, representation learning, graphical models)
  - The class labels (directly related to the task/problem) of training data is unknown
  - Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data
- Semi-supervised Learning
  - AKA. Weakly supervised
  - A small amount of labeled data with a large amount of unlabeled data during training.
  - Active Learning – Human in the loop

Classification vs. Prediction

- Classification
  - predicts categorical class labels (discrete or nominal)
  - classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data
- Prediction
  - models continuous-valued functions, i.e., predicts unknown or missing values
- Typical applications
  - Credit/loan approval
  - Medical diagnosis: if a tumor is cancerous or not
  - Fraud detection: if a transaction is fraudulent
  - Web page categorization: which category it is

Generative vs Discriminative Model

- Generative Model
  - Separately model $P(x|y)$ and $P(y)$ from the training set. Classifying using the Bayesian rule:

$$P(y = 1|x) = \frac{P(x|y = 1)\, p(y = 1)}{p(x)}$$

  - Generally converge faster (efficient in training) and requires less data.
  - Naïve Bayesian, Linear discriminant analysis, …

- Discriminative Model
  - Direct estimation of $P(y|x)$
  - Requires more data but has lower asymptotic error (performs better).
  - Decision Tree, SVM, Logistic Regression, K-NN, …

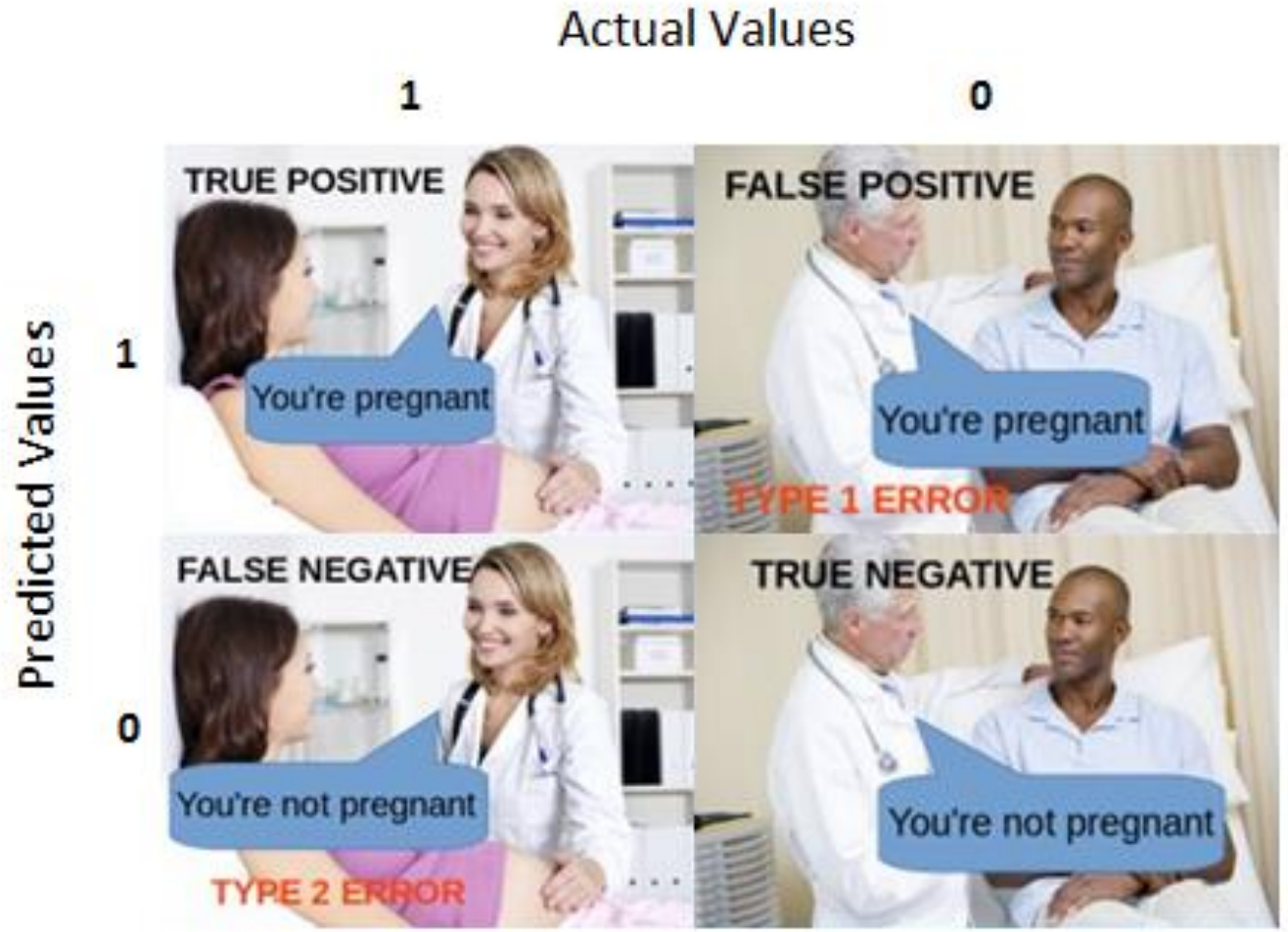
Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems* (pp. 841-848).

Performance Metric

- Not all errors are equally important.
- It depends on what you NEED. (Need-driven)

- Classification:
  - TP, FP, TN, FN, Accuracy, Recall, Precision, Sensitivity, Specificity, AUROC, F1, …
- Prediction (Regression)
  - MSE, MAE, MSLE, R2, …

Confusion Matrix, Type I & Type II error



https://medium.com/@neeraj.kumar.iitg/statistical-performance-measures-12bad66694b7

Confusion Matrix

Real class\Predicted class	C_1	C_2
C_1	True positive (TP)	False negative (FN)
C_2	False positive (FP)	True negative (TN)

Fact\Answer	Yes	no	total
yes	6 (TP)	3 (FN)	9 (P)
no	2 (FP)	3 (TN)	5 (N)
total	8	6	14

- Accuracy: percentage of test set tuples that are correctly classified:
  - $Accuracy = \frac{TP+TN}{P+N} = \frac{6+3}{14}$
  - $Recall = \frac{TP}{P} = \frac{6}{9}$ (percentage of positives correctly answered)
  - $Precision = \frac{TP}{TP+FP} = \frac{6}{6+2}$ (chance that a positive answer is correct)

Confusion Matrix – F1

Real class\Predicted class	C_1	C_2
C_1	True positive (TP)	False negative (FN)
C_2	False positive (FP)	True negative (TN)

Fact\Answer	Yes	no	total
yes	6 (TP)	3 (FN)	9 (P)
no	2 (FP)	3 (TN)	5 (N)
total	8	6	14

- F measure: the harmonic mean of precision and recall:
  - $F = \frac{2 * Precision * Recall}{Precision + Recall}$

Generalizability

- How the trained model performs on unseen/future data
  - I.e. how generalizable is the trained model toward future unknown data (actual prediction tasks)

Multi-class Measures

- Micro-average
  - sum up the individual true positives, false positives, and false negatives for different classes and then apply them to get the statistics.

- Macro-average
  - take the average of the chosen metric (e.g. precision, recall etc.) of the system on different classes

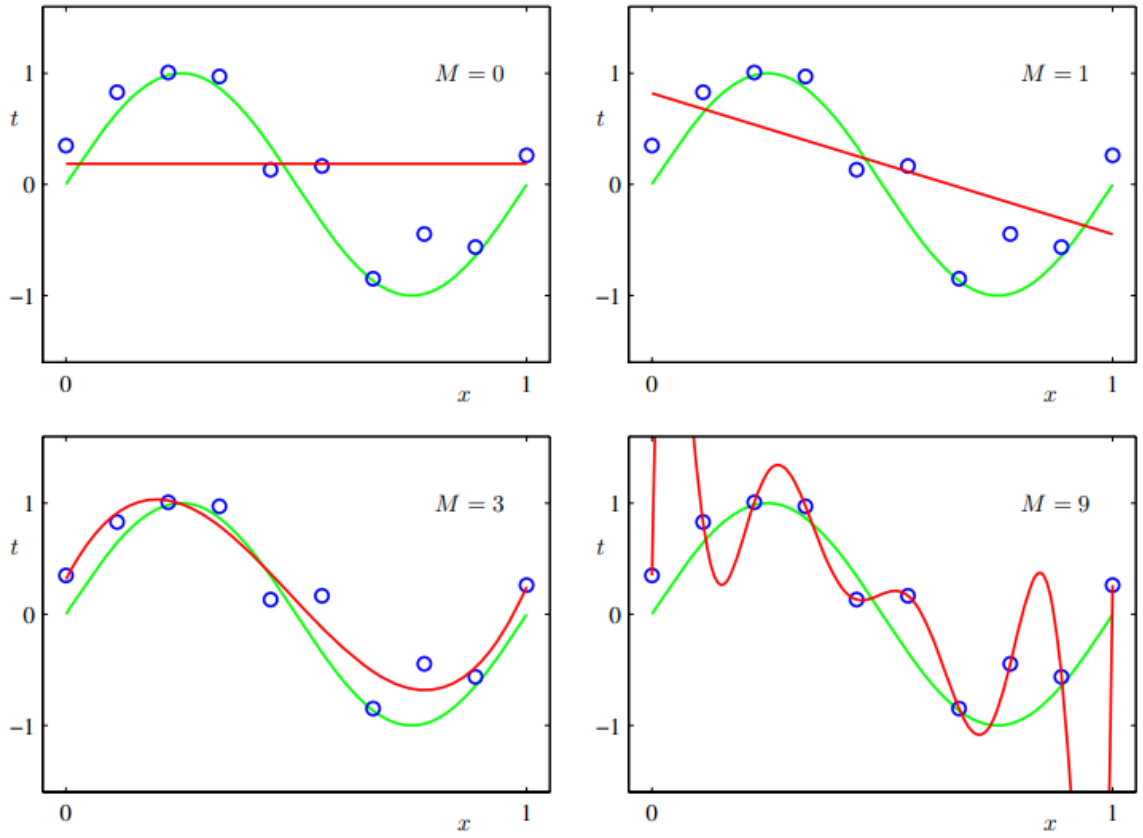
Hold-out method

- Holdout method
  - The given data is randomly partitioned into two independent sets
    - Training set (e.g., 70%) for model construction
    - Test set (e.g., 30%) for accuracy estimation
    - Stratified sampling:
      - Divide the dataset into groups according to the class attribute value.
      - Sample from each group and merge the samples.
      - The resulting dataset share a similar distribution to the original data set w.r.t. the class attribute.
      - Only for the classification problem.

- Random sampling: a variation of holdout
    - Repeat holdout k times, accuracy = avg. of the accuracies obtained

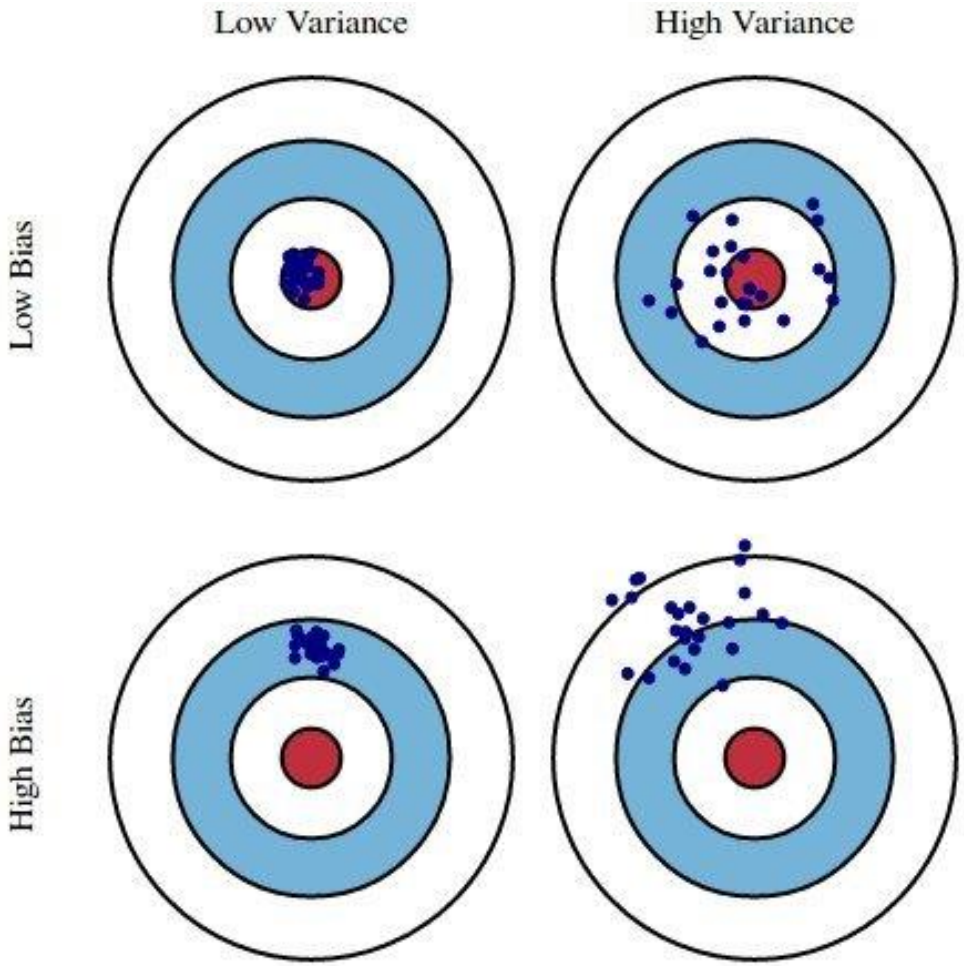
Overfitting

- Adding more degrees of freedom ==> always fit better (to the observations)



Bishop, C. M. (2006). Pattern recognition and machine learning. springer.

Bias vs Variance



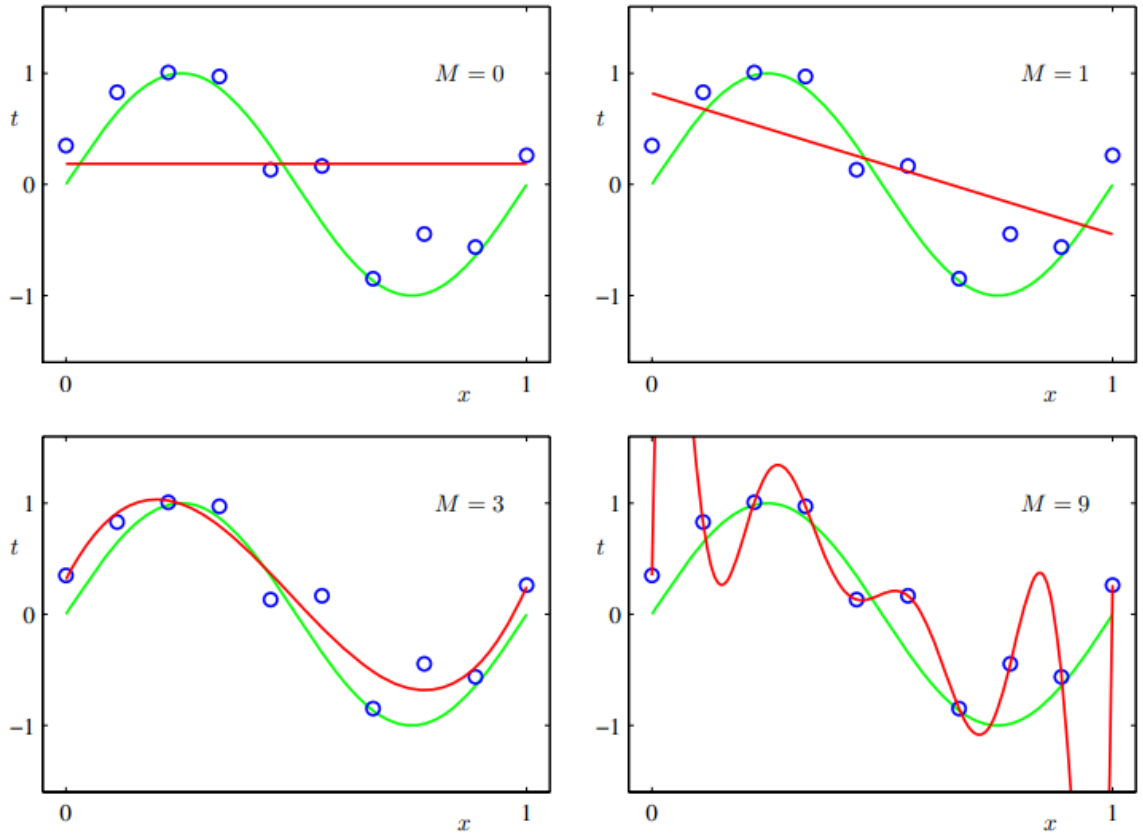
bias is how removed a model's predictions are from correctness,

variance is the degree to which these predictions vary between model iterations.

Understanding the Bias-Variance Tradeoff, by Scott Fortmann-Roe.

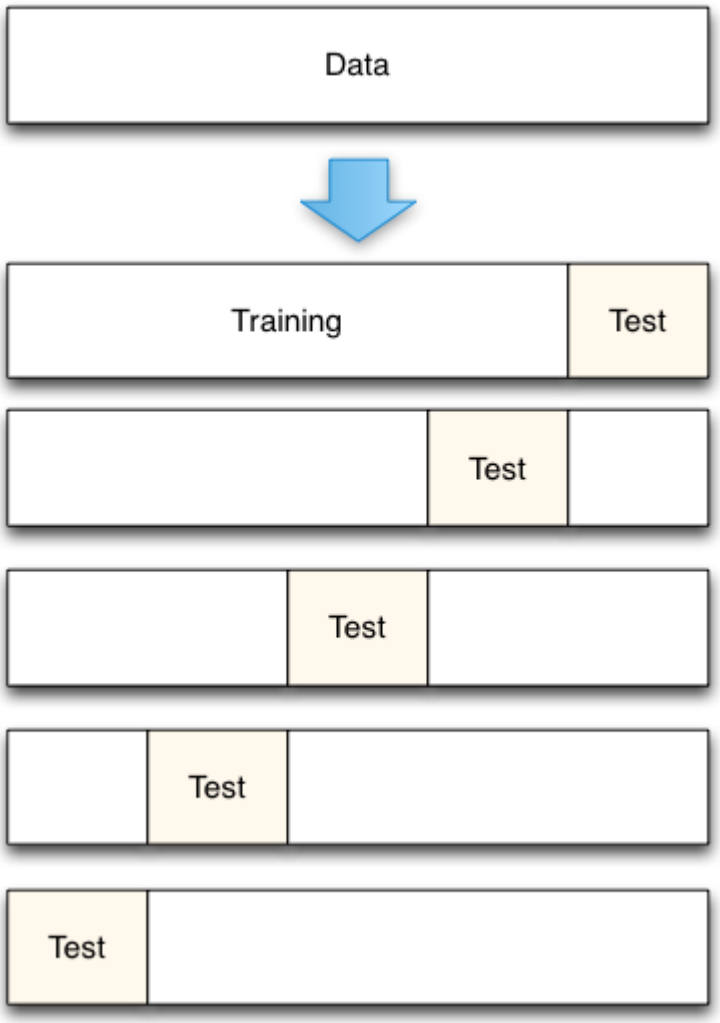
Overfitting – Bias-Variance Point-of-View

- High/Low Bias/Variance?



Bishop, C. M. (2006). Pattern recognition and machine learning. springer.

K-fold cross-validation



Single test-train split (hold-out method): Estimation test error with **high variance**.

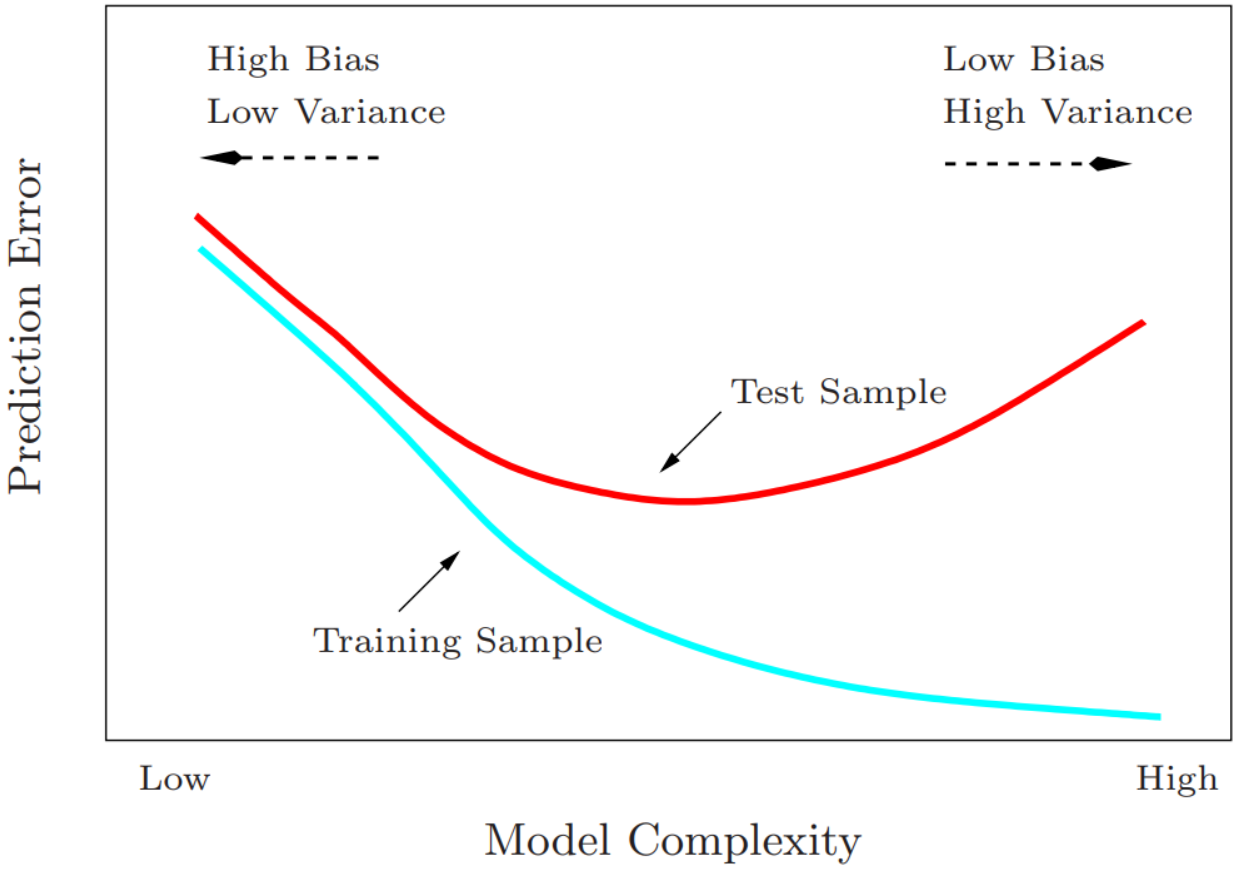
The lower the ***k***, the higher the bias in the error estimates and the less variance.

Conversely, when ***k*** is set equal to the number of instances, the error estimate is then very low in bias but has the possibility of high variance. (Leave-one-out cross-validation)

Cons: computational cost & waste of data

Understanding the Bias-Variance Tradeoff, by Scott Fortmann-Roe.

Biases & Variance Trade-off



Clarke, R., Coladarci, T., & Minium, E. W. (1999). Elements of statistical reasoning. Hoboken, NJ: John Wiley & Sons.

.632 Bootstrapping for error estimation

- Given a data set of ***d*** tuples. The data set is uniformly sampled ***d*** times, with replacement, resulting in a bootstrap sample set of ***d*** samples.
- The bootstrap sample set is used as training set.
- The data tuples that did not make it into the training set end up forming the test set.

.632 Bootstrapping – the resubstitution error

$$Acc(M) = \frac{1}{k}\sum_{i=1}^{k}(0.632 \times Acc(M_i)_{test\_set} + 0.368 \times Acc(M_i)_{train\_set}),$$

1. Each tuple has a probability of 1/d of being selected,
2. so the probability of not being chosen is (1 − 1/d).
3. We have to select d times, so the probability that a tuple will not be chosen during this whole time is (1 − 1/d) d .
4. If d is large, the probability approaches e −1 = 0.368.7
5. Thus, 36.8% of tuples will not be selected for training and thereby end up in the test set, and the remaining 63.2% will form the training set.

Efron, B., & Tibshirani, R. (1997). Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, *92*(438), 548-560.

Hyper-parameter Optimization

- Classifier/Estimator has several configurable parameters
  - SVM: kernel type, C value, gamma value, etc.
  - Decision Tree: maximum depth, pruning conditions, etc.
  - Logistic Regression: regularization options, solver types, etc.
  - Selected based on our knowledge, data distribution, and empirical evaluation. (A process of educated guess and trial-and-error)

  - Tuning too hard on testing set may overfit the testing set.

Improved Hold-out Method:

- Validation/Development Set

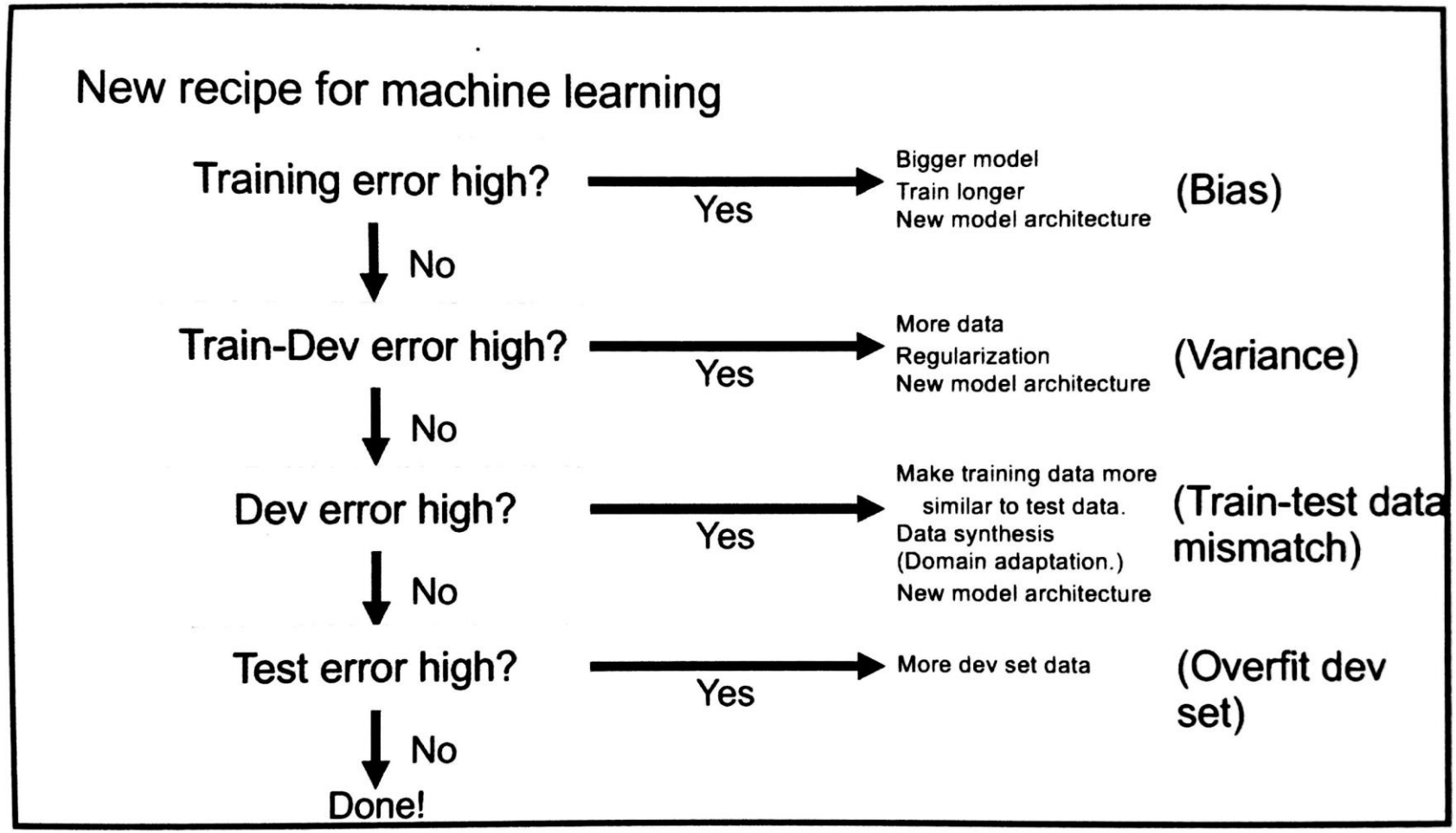
  - Partition the data into three independent sets: training set, validation set, and testing set.

    - Training set: used to train several classifiers based on different parameter configurations.

    - Validation set (1-2): used to evaluate the trained models. Pick the one that achieves the best performance.

    - Testing set: used to evaluate the chosen model and reports its performance. Minimal usage.

Data Science Life Cycle



Andrew Yan-Tak Ng