Lecture 09

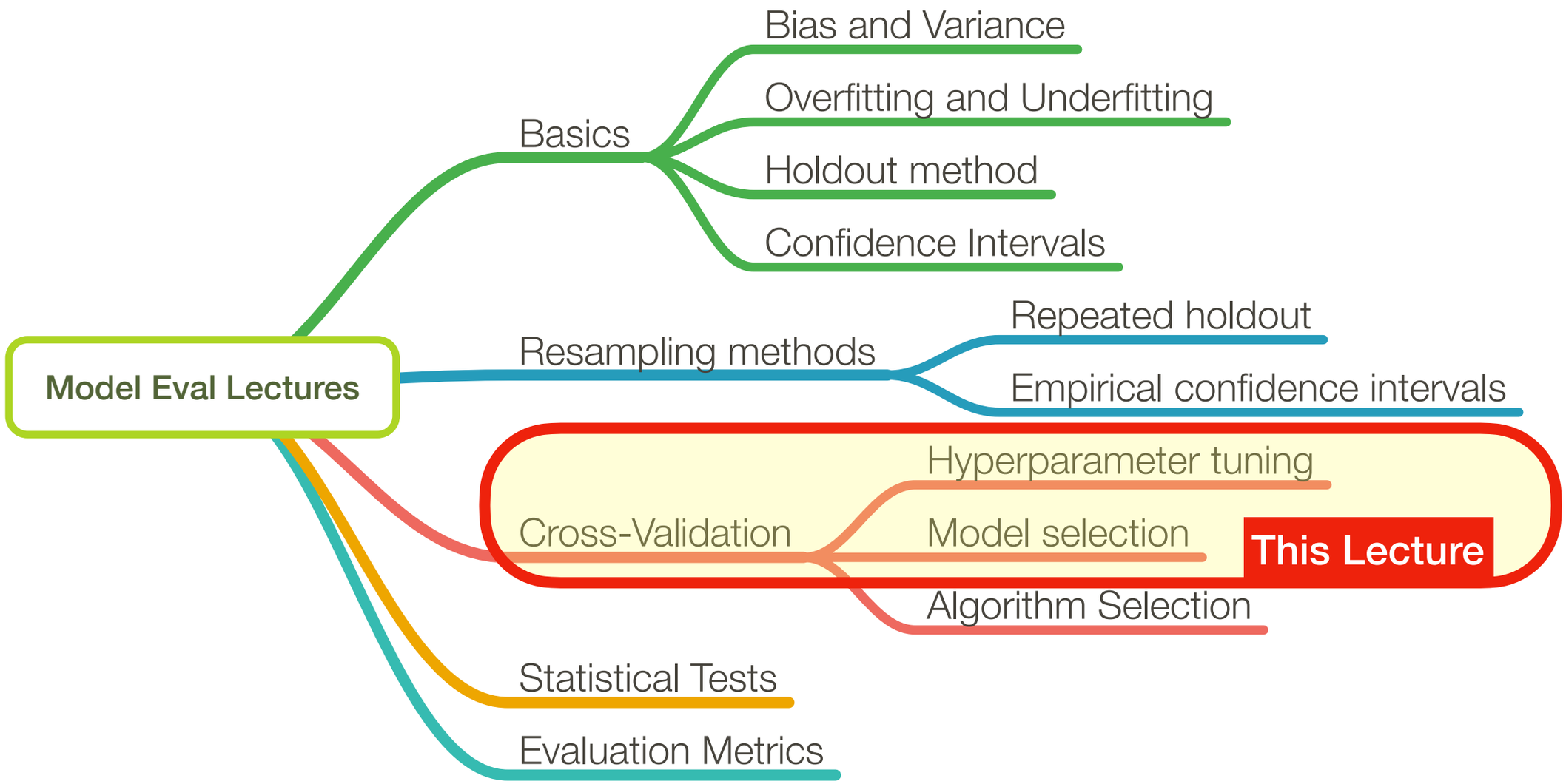
Model Evaluation 2:
Confidence Intervals

STAT 479: Machine Learning, Fall 2018

Sebastian Raschka

http://stat.wisc.edu/~sraschka/teaching/stat479-fs2018/

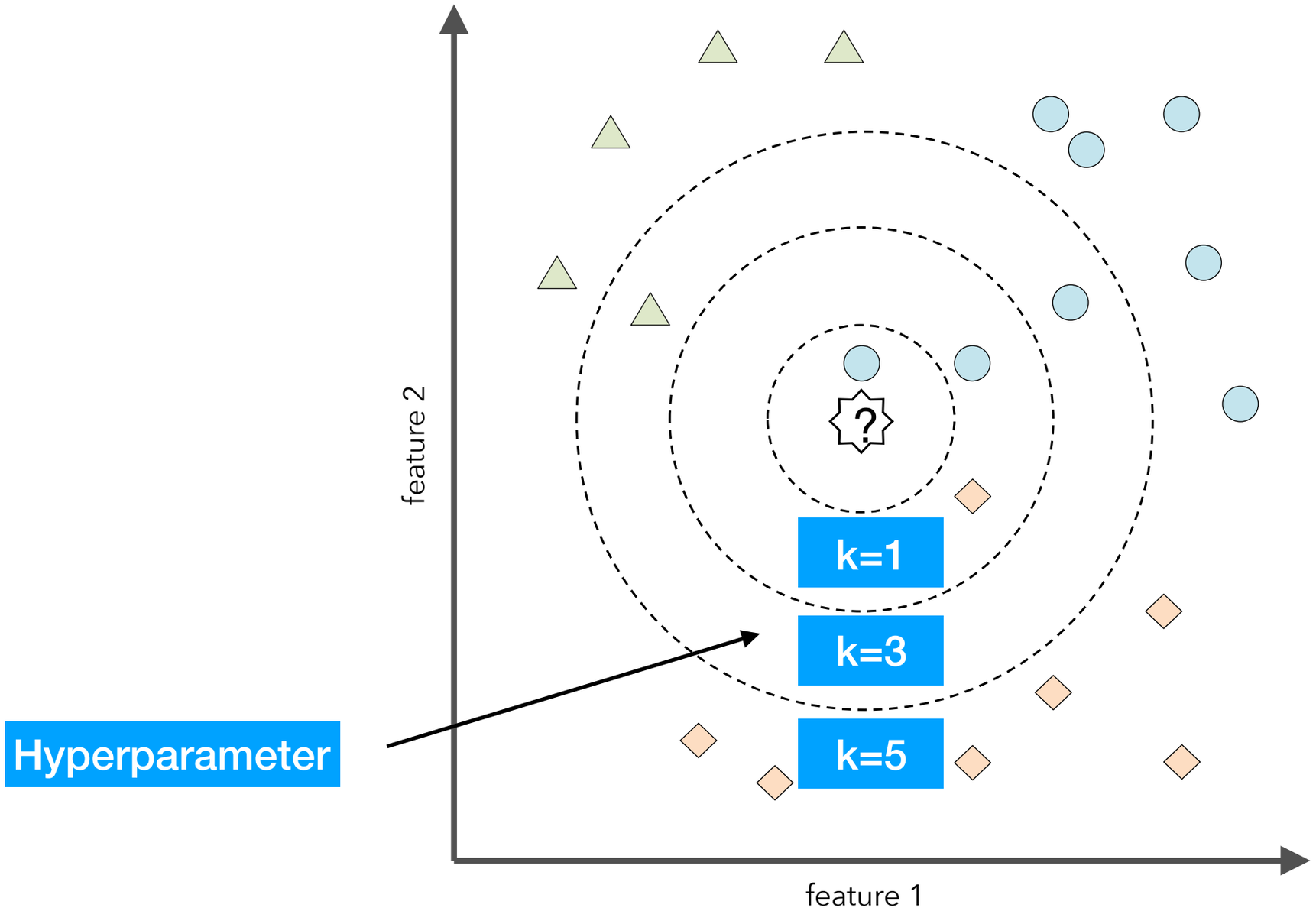
Overview


Model Eval Lectures
Basics
Bias and Variance
Overfitting and Underfitting
Holdout method
Confidence Intervals
Resampling methods
Repeated holdout
Empirical confidence intervals
Cross-Validation
Hyperparameter tuning
Model selection
Algorithm Selection
This Lecture
Statistical Tests
Evaluation Metrics

Hyperparameters

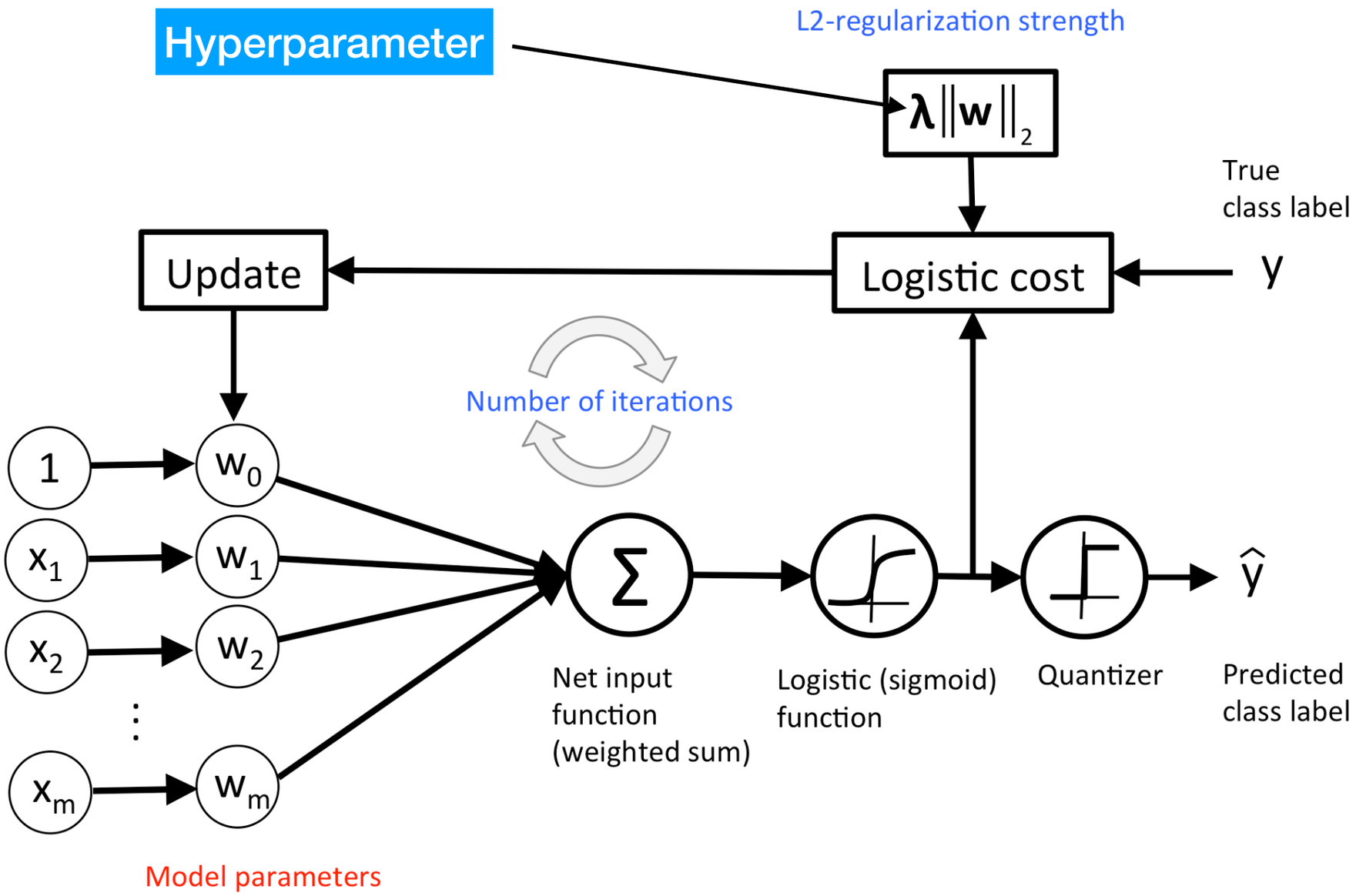
Hyperparameters

nonparametric model: k-nearest neighbors



Hyperparameters

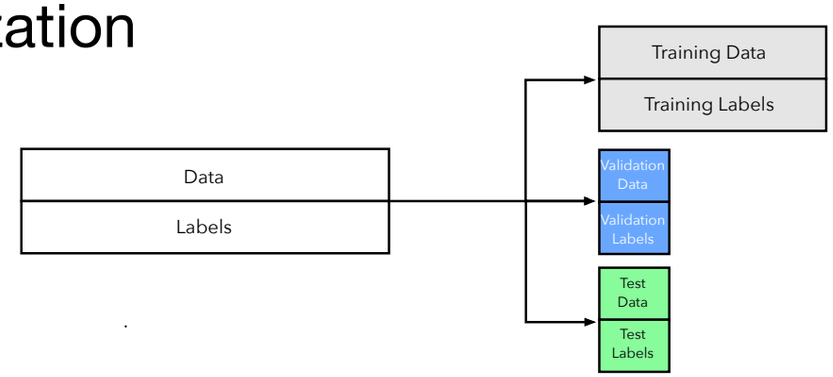
parametric model: logistic regression



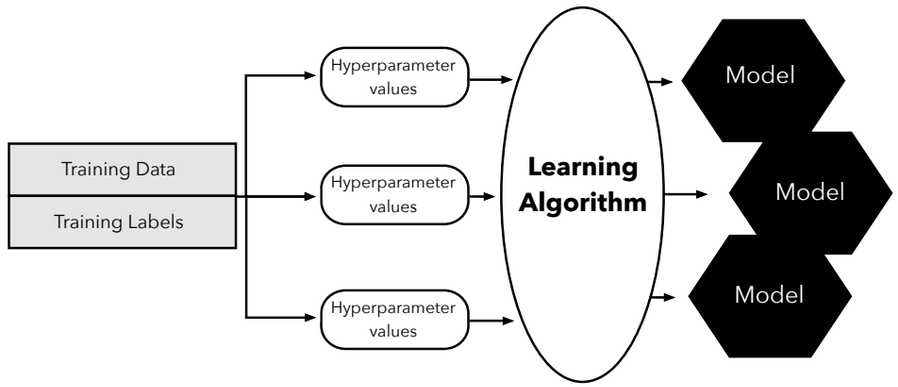
3-Way Holdout

instead of "regular" holdout to avoid "data leakage" during hyperparameter optimization

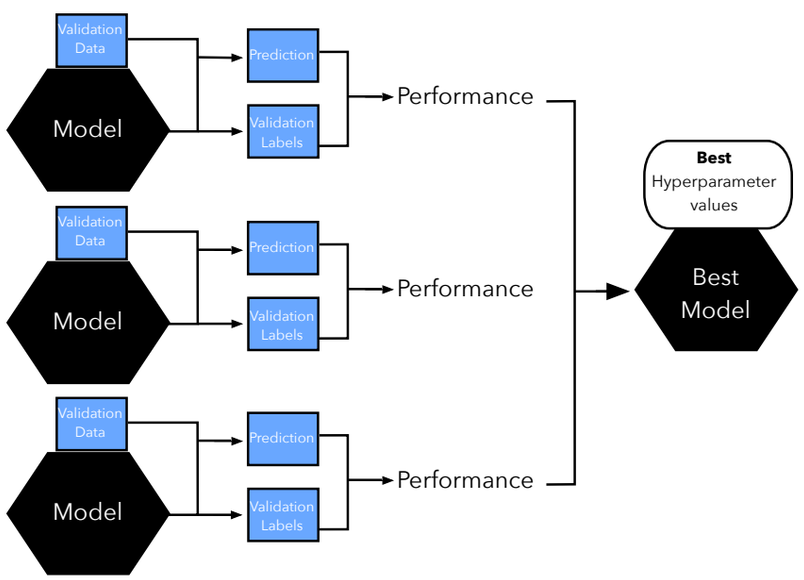
1



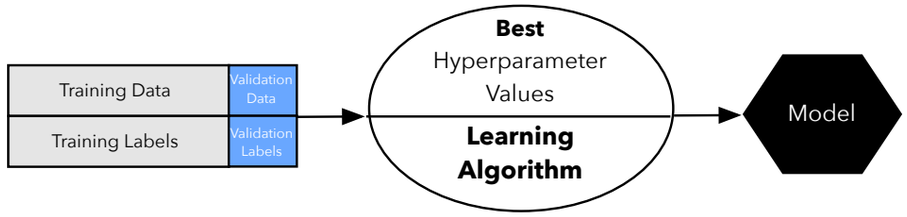
2



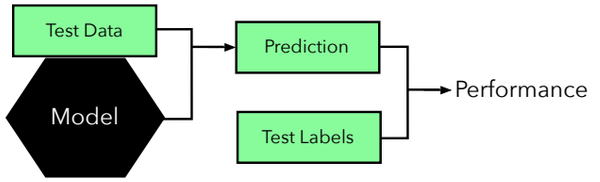
3



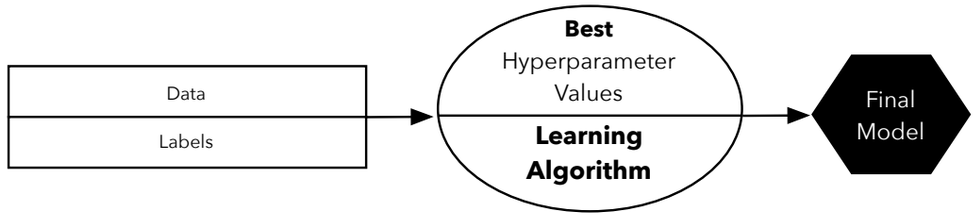
4



5



6



Main points why we evaluate the predictive performance of a model:

1. Want to estimate the generalization performance, the predictive performance of our model on future (unseen) data.

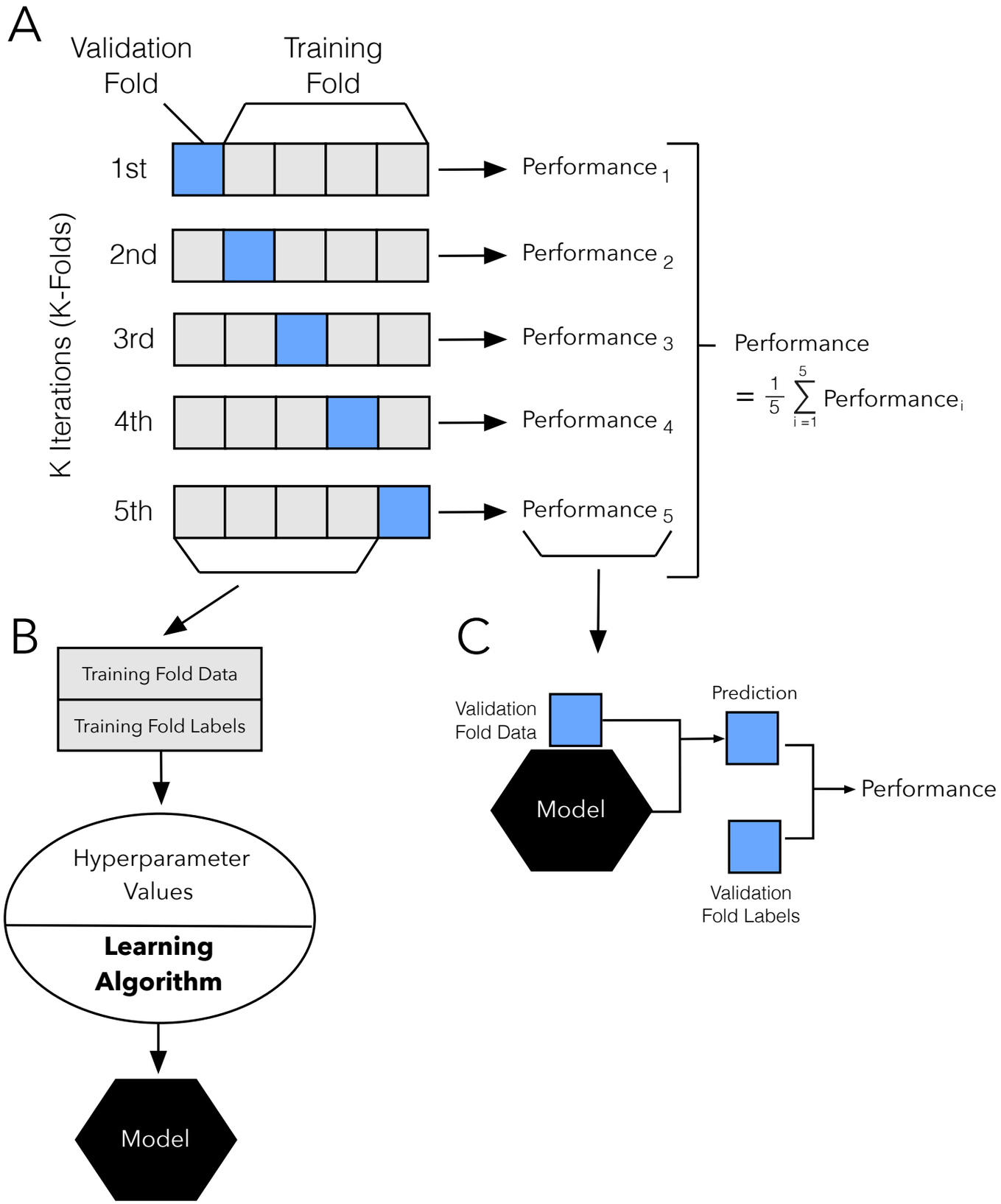
2. Want to increase the predictive performance by tweaking the learning algorithm and selecting the best performing model from a given hypothesis space.

3. Want to identify the ML algorithm that is best-suited for the problem at hand; thus, we want to compare different algorithms, selecting the best-performing one as well as the best performing model from the algorithm's hypothesis space.

k-Fold Cross-Validation Part 1

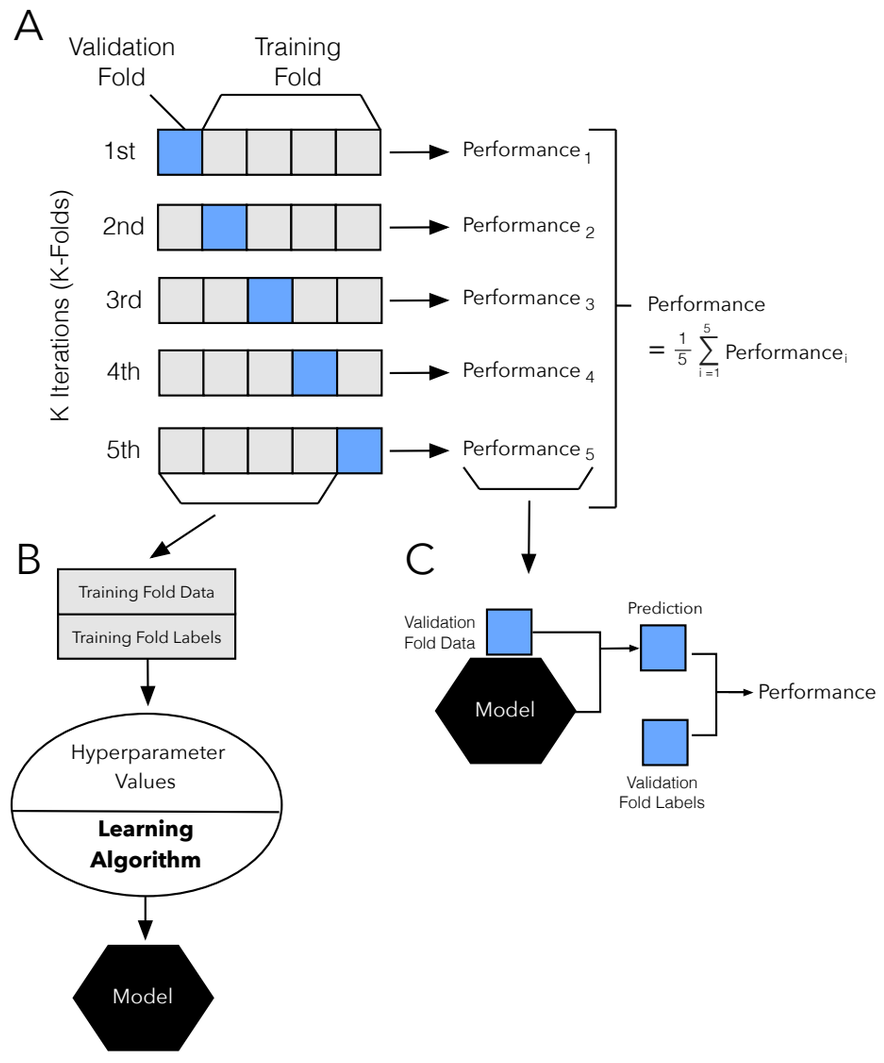
Model Evaluation

k-Fold Cross-Validation

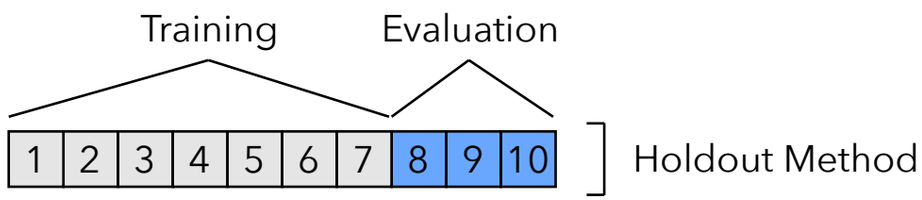


k-Fold Cross-Validation

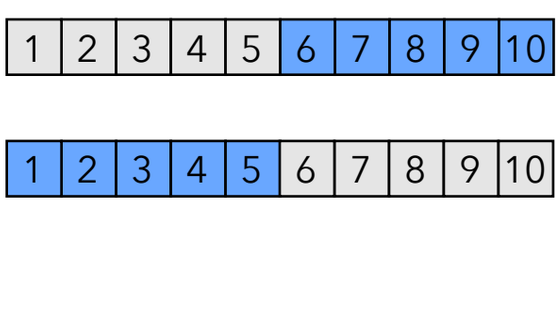
- non-overlapping test folds; utilizes all data for testing
- overlapping training folds
- some variance estimate from different training sets, (but no unbiased estimate)
- more pessimistic for small k because we withhold data from fitting



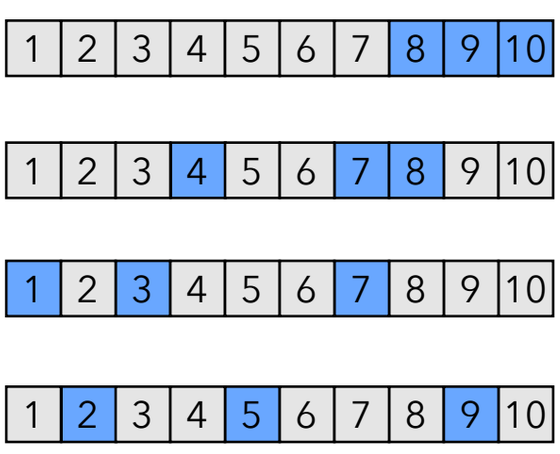
k-Fold CV special cases: k=2 & k=n



Holdout Method

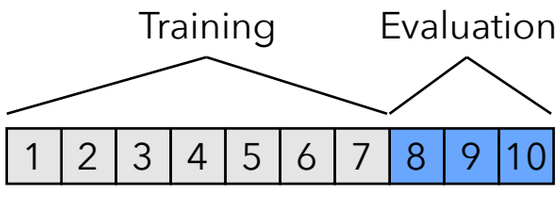


2-Fold Cross-Validation

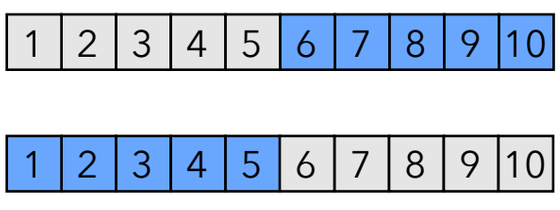


Repeated Holdout

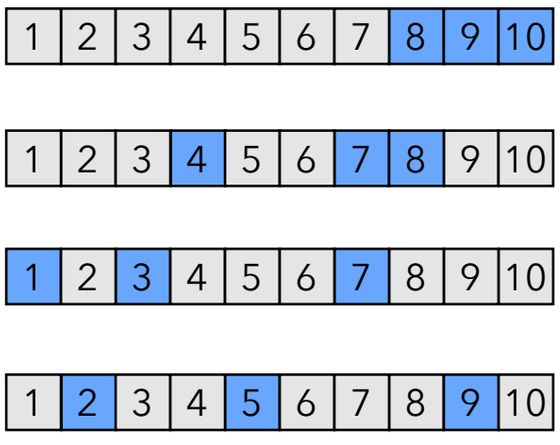
k-Fold CV special cases: k=2 & k=n



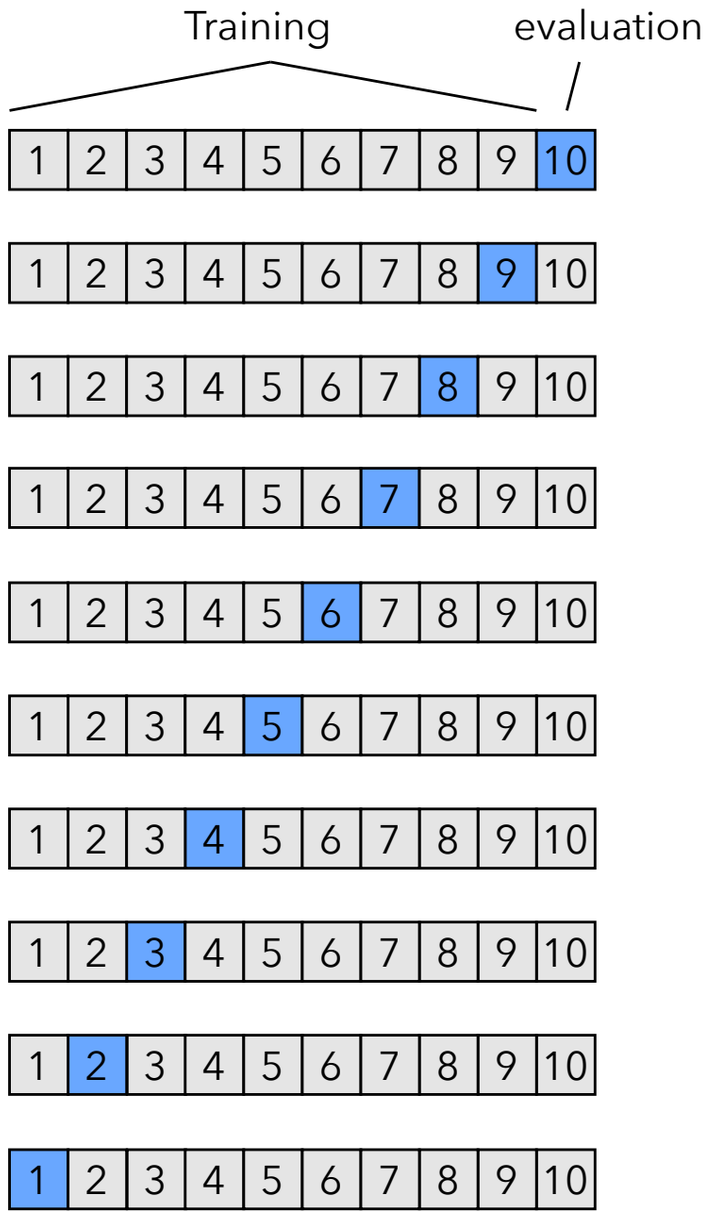
Holdout Method



2-Fold Cross-Validation



Repeated Holdout



k-Fold Cross-Validation

"[...] where available sample sizes are modest, holding back compounds for model testing is ill-advised. This fragmentation of the sample harms the calibration and does not give a trustworthy assessment of fit anyway. It is better to use all data for the calibration step and check the fit by cross-validation, making sure that the cross-validation is carried out correctly. [...] The only motivation to rely on the holdout sample rather than cross-validation would be if there was reason to think the cross-validation not trustworthy -- biased or highly variable. But neither theoretical results nor the empiric results sketched here give any reason to disbelieve the cross-validation results." [1]

1. Hawkins, D. M., Basak, S. C., & Mills, D. (2003). Assessing model fit by cross-validation. *Journal of chemical information and computer sciences*, *43*(2), 579-586.

LOOCV vs Holdout

Experiment	**Mean**	**Standard deviation**
True R^2 — q^2	0.010	0.149
True R^2 — hold 50	0.028	0.184
True R^2 — hold 20	0.055	0.305
True R^2 — hold 10	0.123	0.504

1. Hawkins, D. M., Basak, S. C., & Mills, D. (2003). Assessing model fit by cross-validation. *Journal of chemical information and computer sciences*, *43*(2), 579-586.

The reported "mean" refers to the averaged difference between the true coefficients of determination (R2) and the coefficients obtained via LOOCV (here called q^2) after repeating this procedure on different 100-example training

LOOCV vs Holdout

Experiment	**Mean**	**Standard deviation**
True R^2 — q^2	0.010	0.149
True R^2 — hold 50	0.028	0.184
True R^2 — hold 20	0.055	0.305
True R^2 — hold 10	0.123	0.504

1. Hawkins, D. M., Basak, S. C., & Mills, D. (2003). Assessing model fit by cross-validation. *Journal of chemical information and computer sciences*, *43*(2), 579-586.

The reported "mean" refers to the averaged difference between the true coefficients of determination (R2) and the coefficients obtained via LOOCV (here called q^2) after repeating this procedure on different 100-example training

In rows 2-4, the researchers used the holdout method for fitting models to the 100-example training sets, and they evaluated the performances on holdout sets of sizes 10, 20, and 50 samples. Each experiment was repeated 75 times, and the mean column shows the average difference between the estimated R^2 and the true R^2 values.

Problems with LOOCV for Classification

- While LOOCV is almost unbiased, one downside of using LOOCV over k-fold cross-validation with k < n is the large variance of the LOOCV estimate.

- LOOCV is "defect" when using a discontinuous loss-function such as the 0-1 loss in classification or even in continuous loss functions such as the mean-squared-error.

- LOOCV has high variance because the test set only contains one sample

Problems with LOOCV for Classification

"With *k*=*n*, the cross-validation estimator is approximately unbiased for the true (expected) prediction error, but can have high variance because the n "training sets" are so similar to one another." [1]

[1] Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1, No. 10). New York, NY, USA:: Springer series in statistics.

Problems with LOOCV for Classification

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \text{Cov}(X_i, X_j) = \sum_{i=1}^{n} \text{Var}(X_i) + 2 \sum_{1 \leq i < j \leq n} \text{Cov}(X_i, X_j)$$

For correlated variables, the variance of their sum is the sum of their covariances

Or in other words, we can attribute the high variance to the fact that the mean of highly correlated variables has a higher variance than the mean of variables that are not highly correlated?

Empirical Study and Recommendation

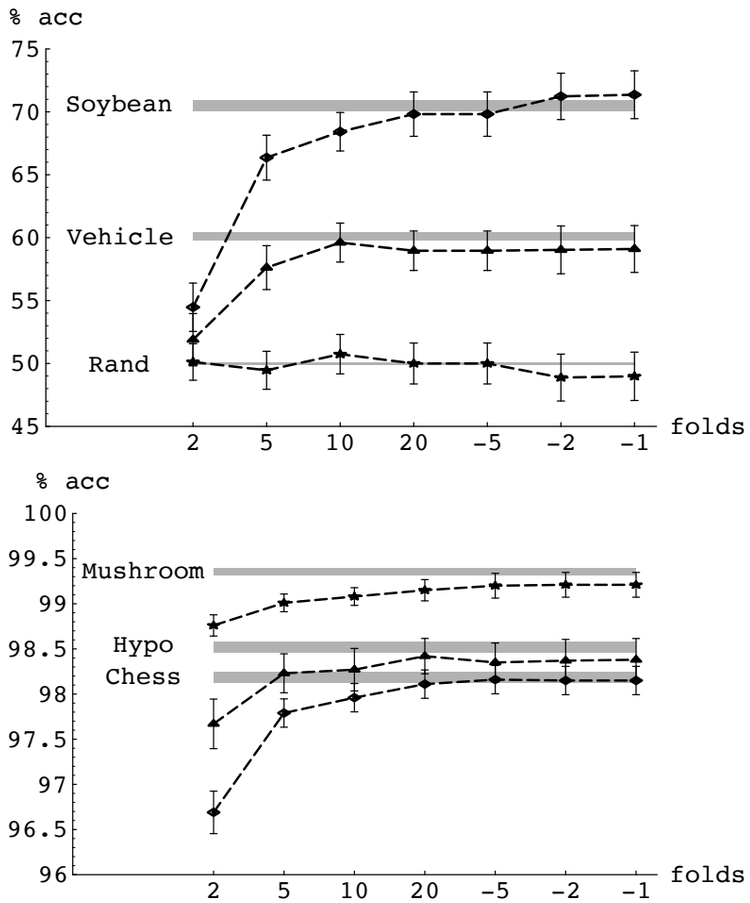


Figure 1: C4.5: The bias of cross-validation with varying folds. A negative k folds stands for leave-k-out. Error bars are 95% confidence intervals for the mean. The gray regions indicate 95% confidence intervals for the true accuracies. Note the different ranges for the accuracy axis.

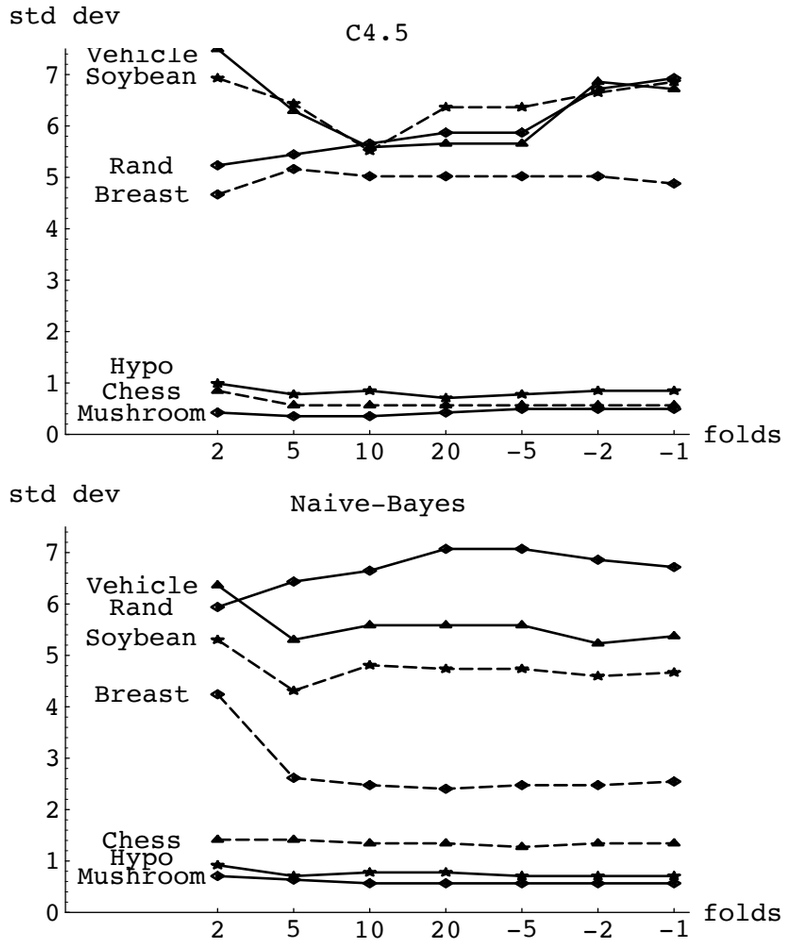


Figure 3: Cross-validation: standard deviation of accuracy (population). Different line styles are used to help differentiate between curves.

Summarizing k-Fold CV for Model Evaluation

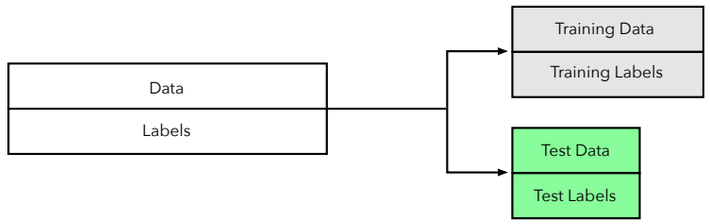
What happens if we increase *k*?

- The bias of the performance estimator decreases (more accurate)
- The variance of the performance estimators increases (more variability)
- The computational cost increases (more iterations, larger training sets during fitting)
- Exception: decreasing the value of k in k-fold cross-validation to small values (for example, 2 or 3) also increases the variance on small datasets due to random sampling effects.

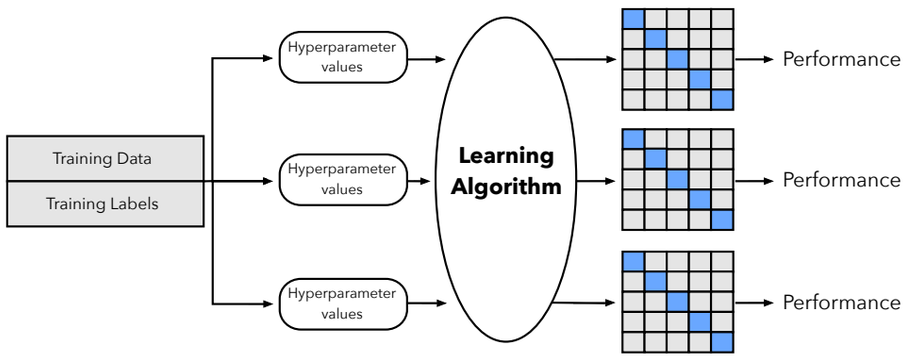
k-Fold Cross-Validation Part 2

Model Selection

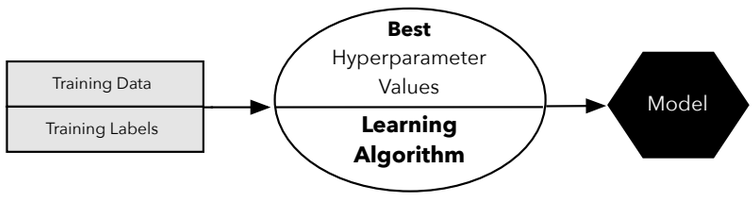
1


Data
Labels
Training Data
Training Labels
Test Data
Test Labels

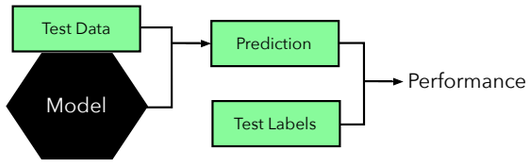
2


Training Data
Training Labels
Hyperparameter values
Hyperparameter values
Hyperparameter values
Learning Algorithm
Performance
Performance
Performance

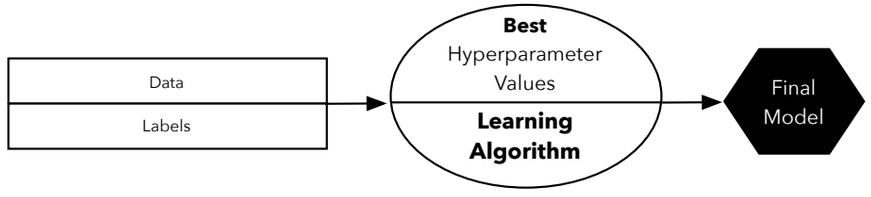
3


Training Data
Training Labels
Best Hyperparameter Values
Learning Algorithm
Model

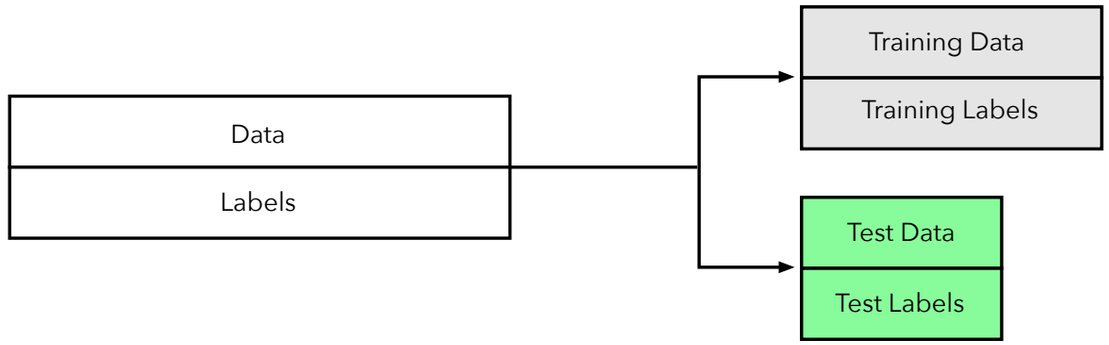
4


Test Data
Model
Prediction
Test Labels
Performance

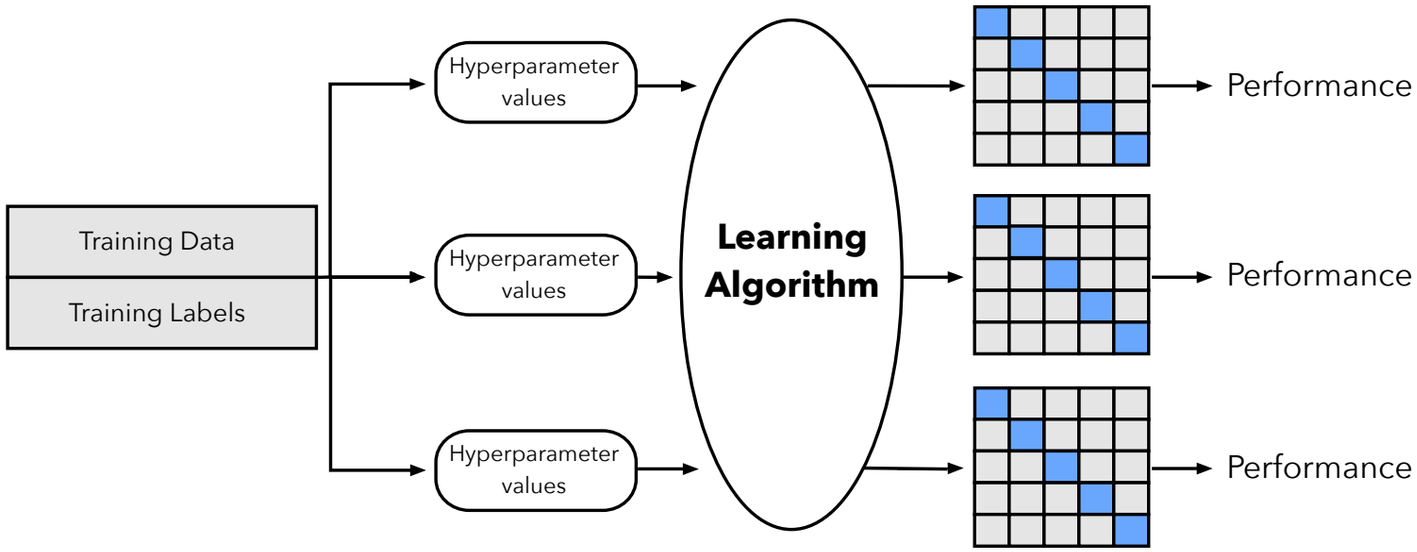
5


Data
Labels
Best Hyperparameter Values
Learning Algorithm
Final Model

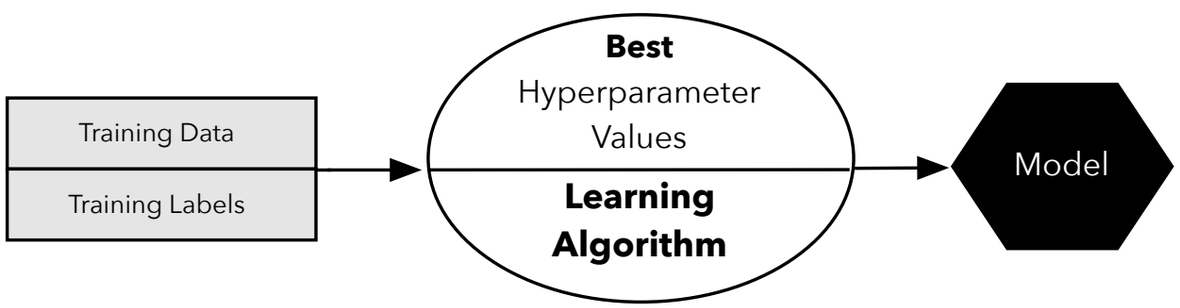
1


Data
Labels
Training Data
Training Labels
Test Data
Test Labels

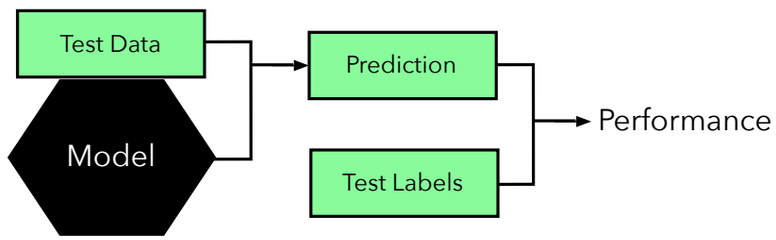
2


Training Data
Training Labels
Hyperparameter values
Hyperparameter values
Learning Algorithm
Performance
Performance

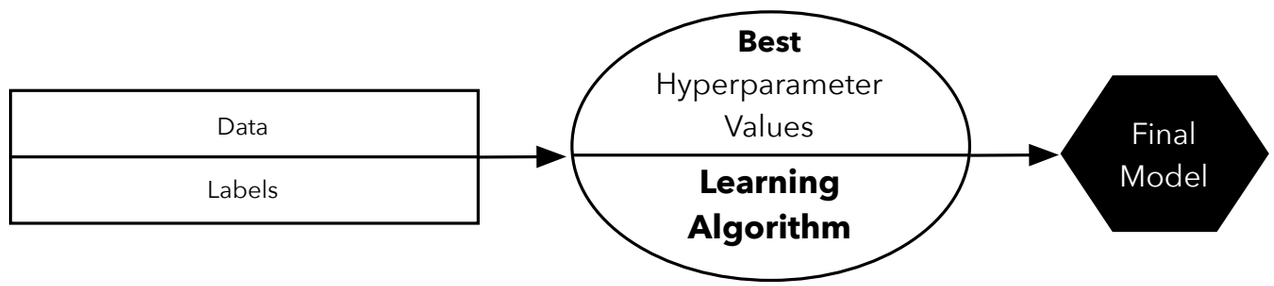
3


Training Data
Training Labels
Best Hyperparameter Values
Learning Algorithm
Model

4


Test Data
Model
Prediction
Test Labels
Performance

5


Data
Labels
Best
Hyperparameter
Values
Learning
Algorithm
Final
Model

The Law of Parsimony

Occam's Razor: "Among competing hypotheses, the one with the fewest assumptions should be selected."

The Law of Parsimony

"Simpler models are more accurate. This belief is sometimes equated with Occam's razor, but the razor only says that simpler explanations are preferable, not why. They're preferable because they're easier to understand, remember, and reason with. Sometimes the simplest hypothesis consistent with the data is less accurate for prediction than a more complicated one. Some of the most powerful learning algorithms output models that seem gratuitously elaborate -- sometimes even continuing to add to them after they've perfectly fit the data -- but that's how they beat the less powerful ones."

Pedro Domingos: "Then Myths about Machine Learning"
https://medium.com/@pedromdd/ten-myths-about-machine-learning-d888b48334a3

The 1-standard error method

... However, if two models perform equally well, the simpler one seems more likely (among other advantages)

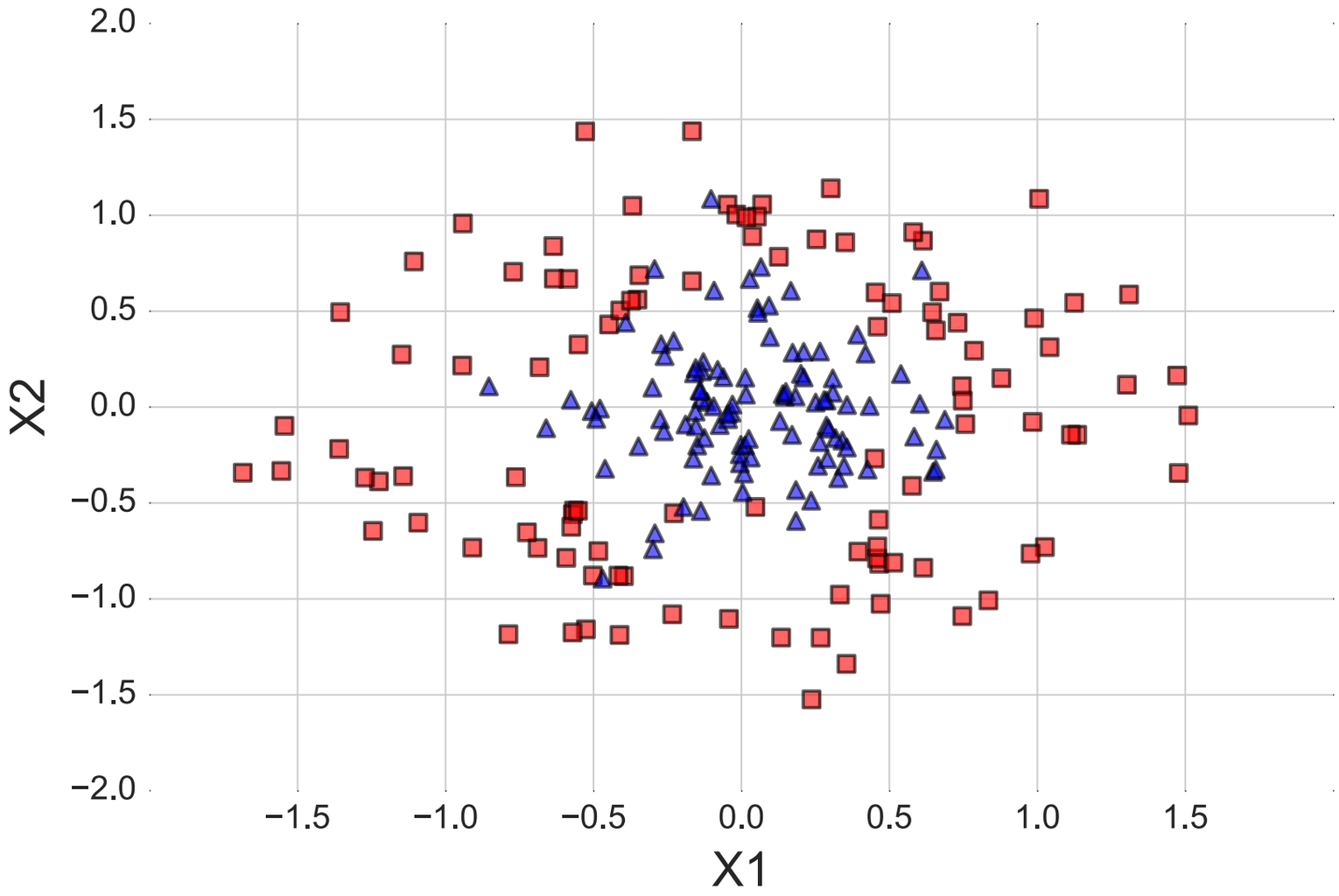
The 1-standard error method

... However, if two models perform equally well, the simpler one seems more likely (among other advantages)

1. Consider the numerically optimal estimate and its standard error.

2. Select the model whose performance is within one standard error of the value obtained in step 1.

The 1-standard error method



(Some toy data I generated via scikit-learn)

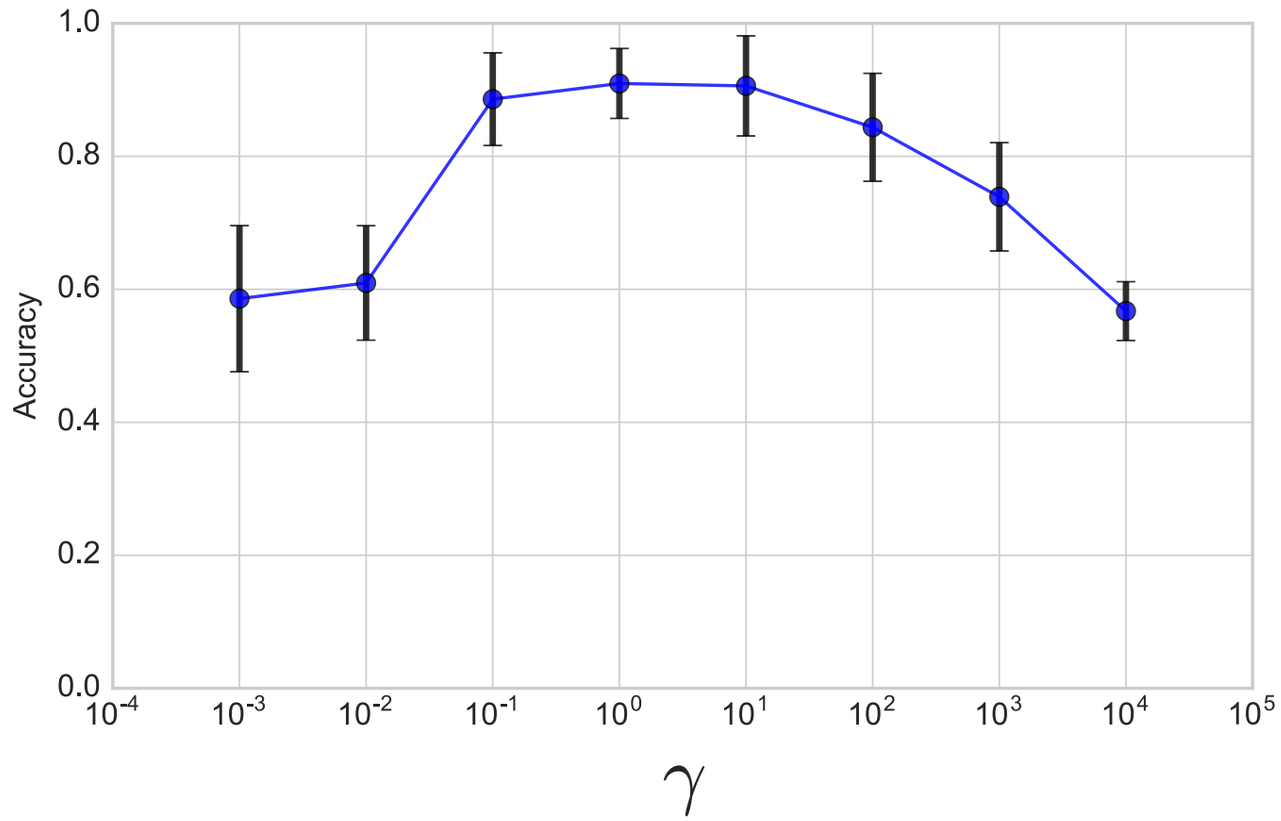
The 1-standard error method

Consider a RBF-kernel SVM, where gamma controls the influence of the training points (don't need to know the details, yet)

Gaussian/RBF-kernel: $K(x_i, x_j) = exp(-\gamma ||x_i - x_j||^2), \gamma > 0.$

The 1-standard error method

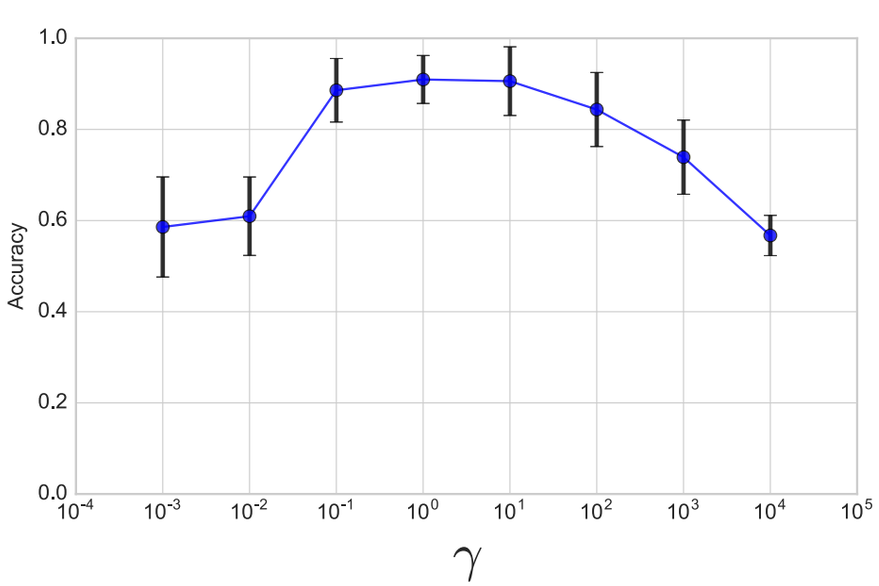
Which parameter would you select?

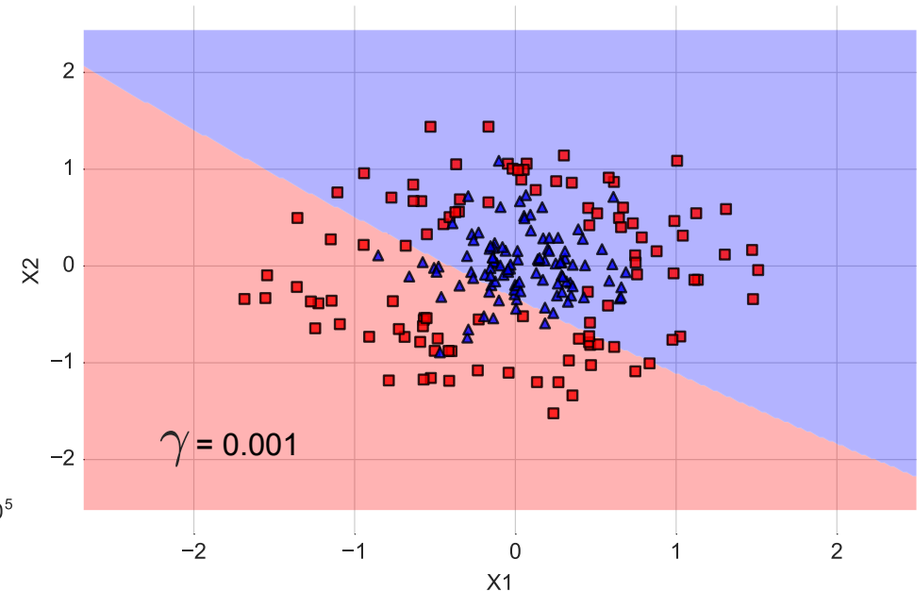


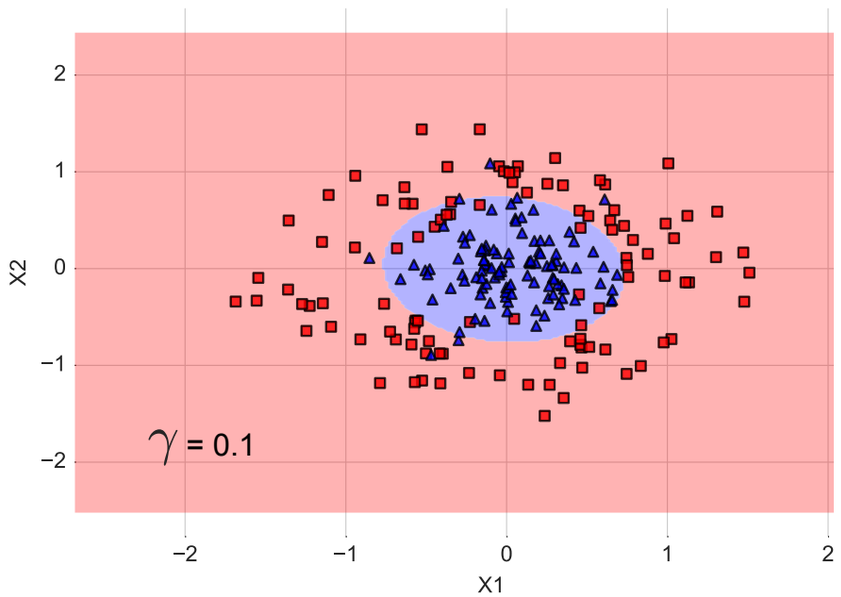
(note: here I used 10-fold CV)

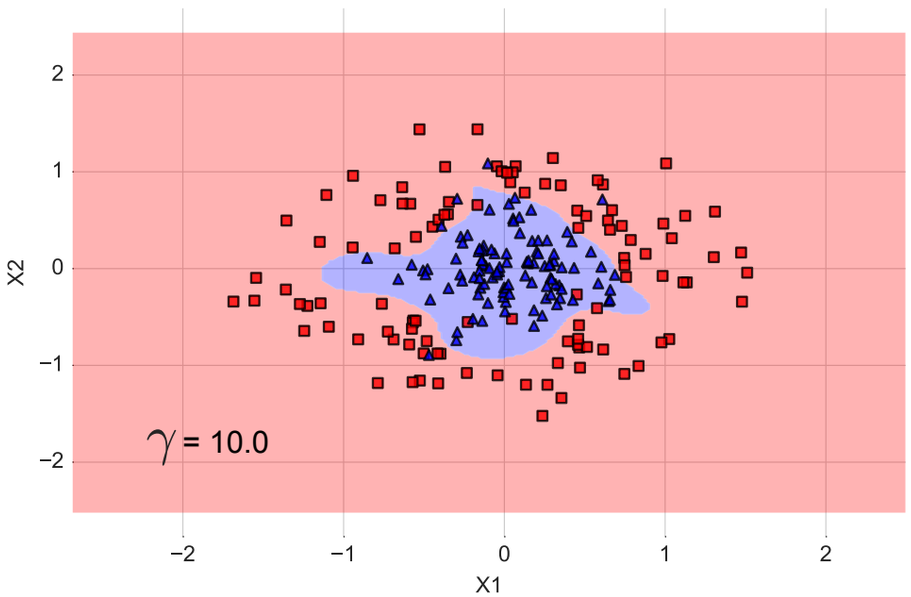
The 1-standard error method

Which parameter would you select?









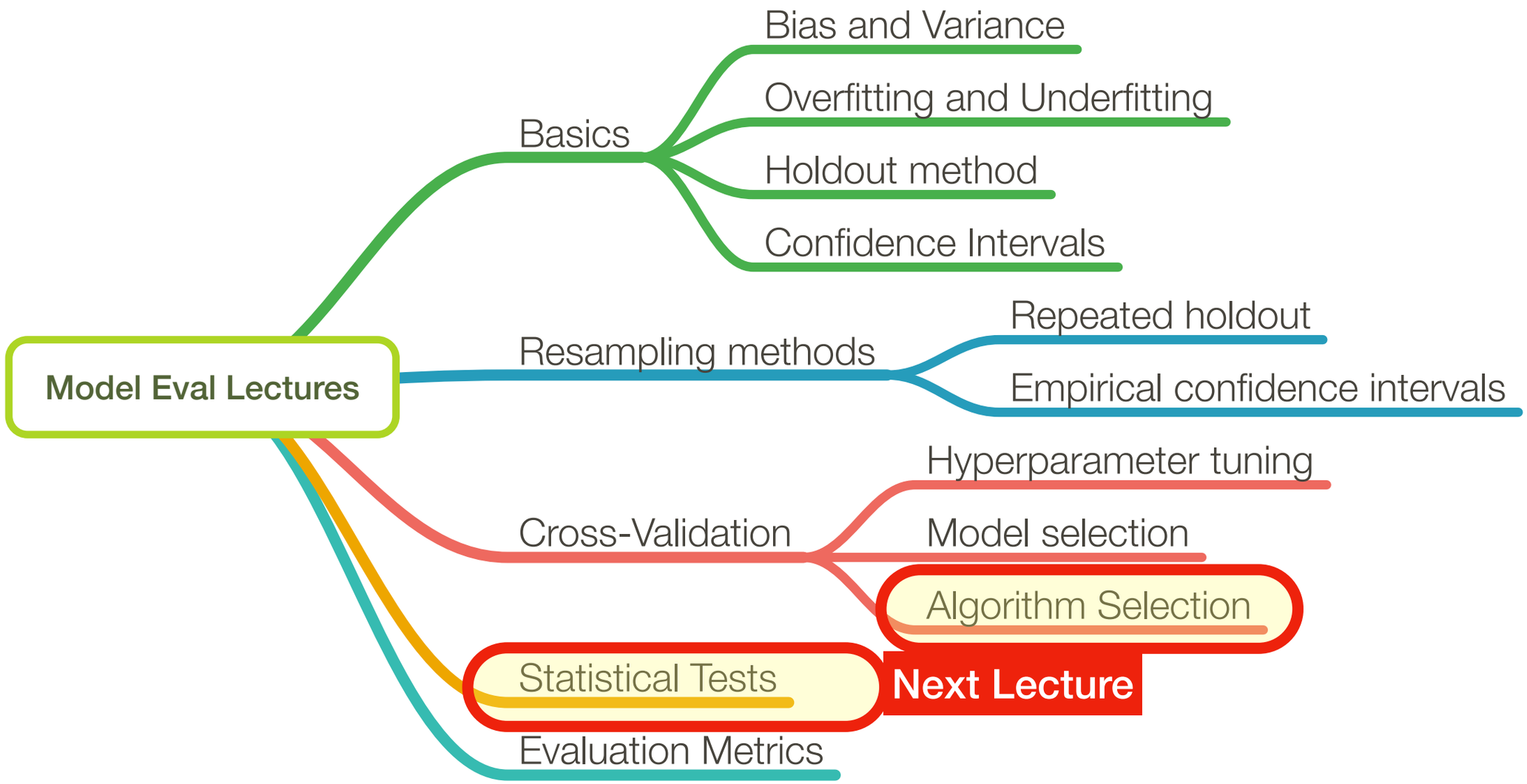
(note: here I used 10-fold CV)

Nested Cross-Validation for Algorithm Selection

Code Examples

https://github.com/rasbt/stat479-machine-learning-fs18/blob/master/10_eval-cv/10_eval-cv_code.ipynb

Overview


Model Eval Lectures
Basics
Bias and Variance
Overfitting and Underfitting
Holdout method
Confidence Intervals
Resampling methods
Repeated holdout
Empirical confidence intervals
Cross-Validation
Hyperparameter tuning
Model selection
Algorithm Selection
Statistical Tests
Next Lecture
Evaluation Metrics