

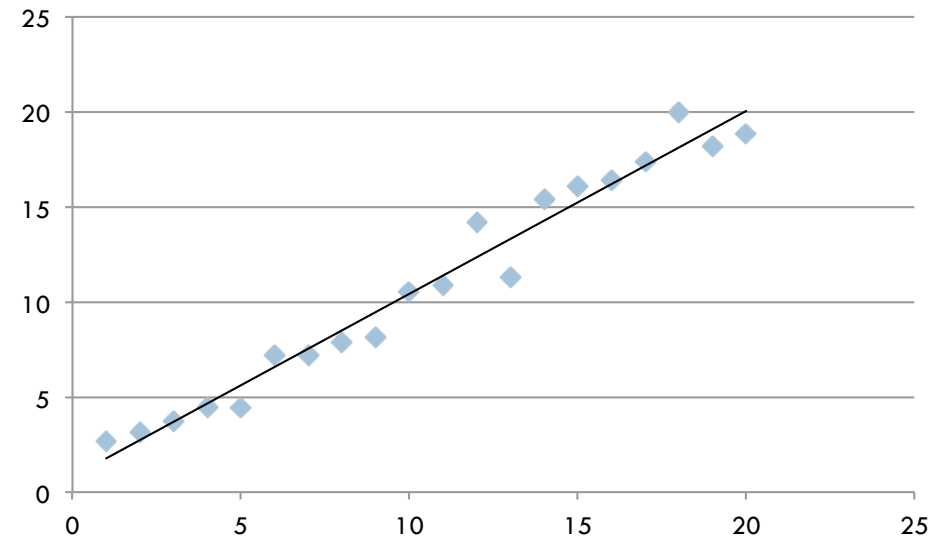
Week 2 Video 5

Cross-Validation and Over-Fitting

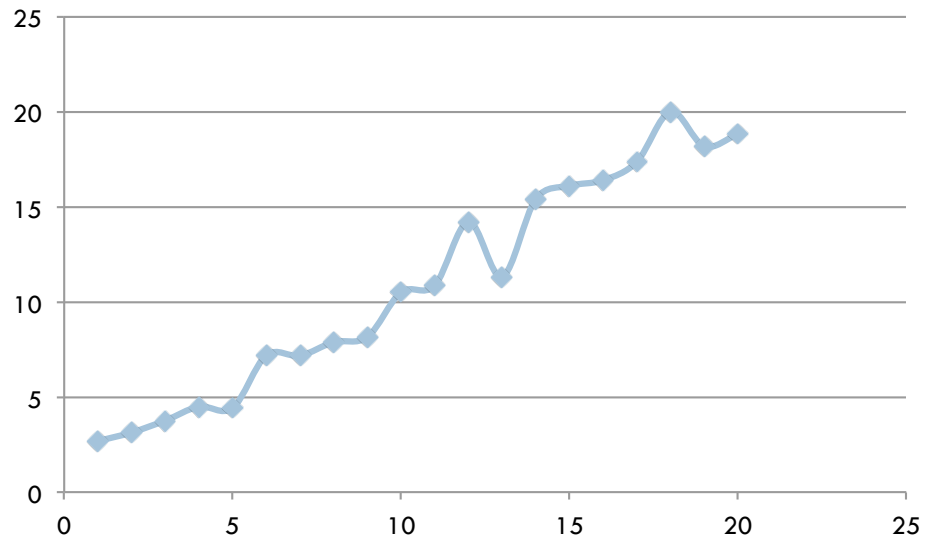
Over-Fitting

- I've mentioned over-fitting a few times during the last few weeks
- Fitting to the noise as well as the signal

Over-Fitting



Good fit



Over fit

Reducing Over-Fitting

- Use simpler models
 - ▣ Fewer variables (BiC, AIC, Occam's Razor)
 - ▣ Less complex functions (MDL)

Eliminating Over-Fitting?

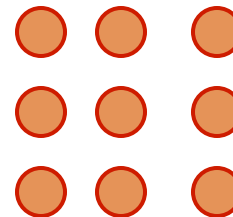
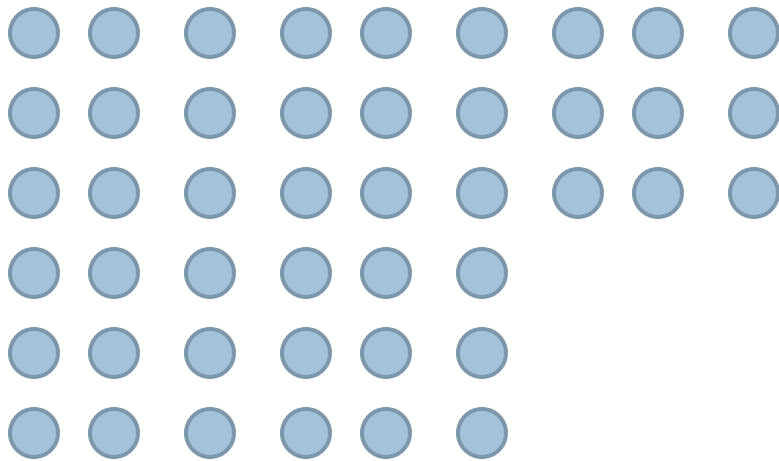
- Every model is over-fit in some fashion
- The questions are:
 - ▣ How bad?
 - ▣ What is it over-fit to?

Assessing Generalizability

- Does your model transfer to new contexts?
- Or is it over-fit to a specific context?

Training Set/Test Set

- Split your data into a training set and test set



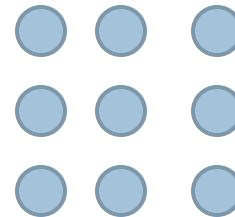
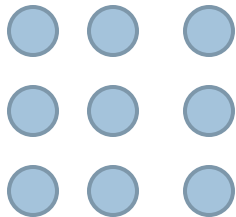
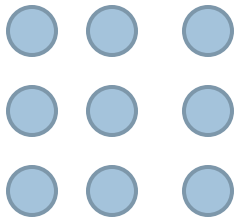
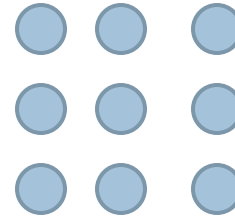
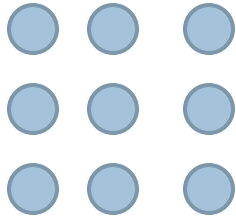
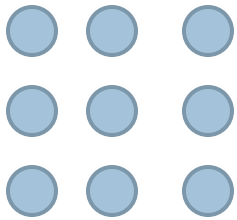
Notes

- Model tested on unseen data
- But uses data unevenly

Cross-validation

9

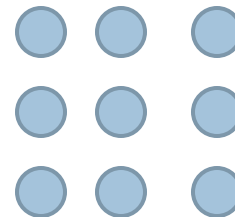
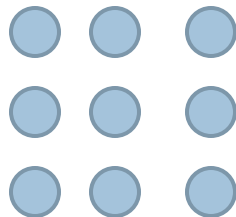
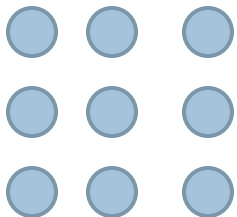
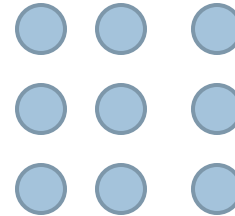
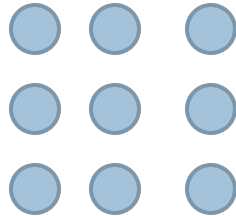
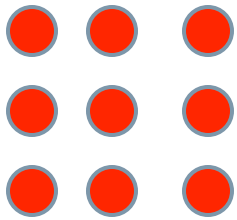
- Split data points into N equal-size groups



Cross-validation

10

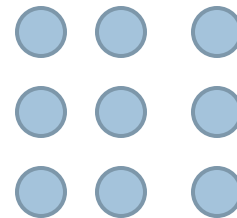
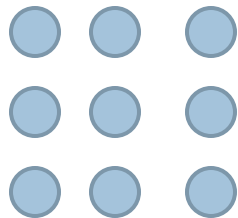
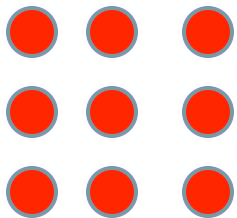
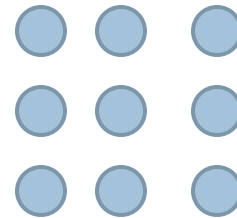
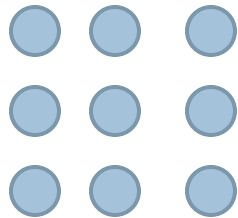
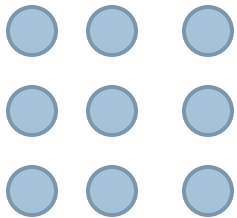
- Train on all groups but one, test on last group
- For each possible combination



Cross-validation

11

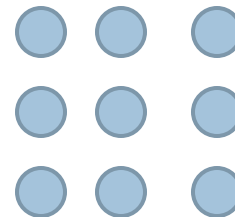
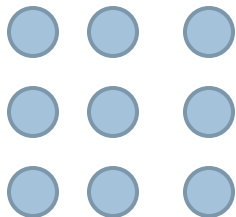
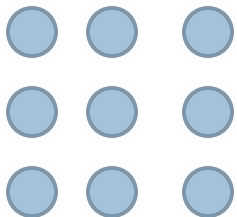
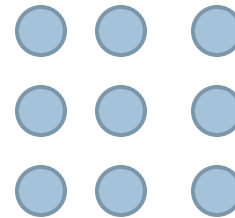
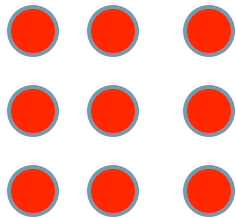
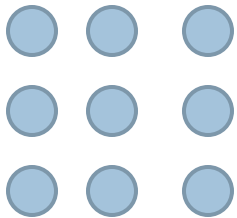
- Train on all groups but one, test on last group
- For each possible combination



Cross-validation

12

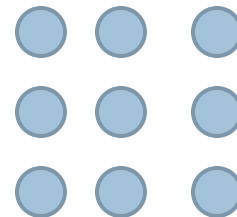
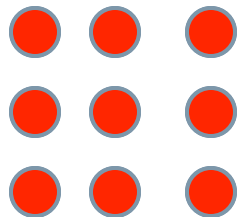
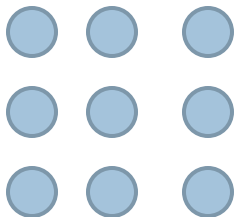
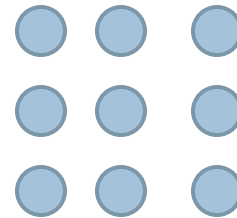
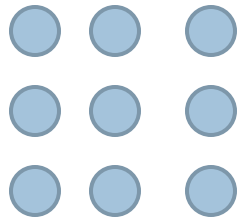
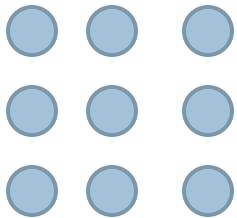
- Train on all groups but one, test on last group
- For each possible combination



Cross-validation

13

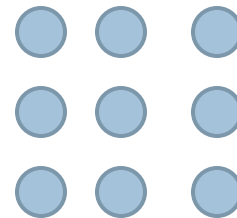
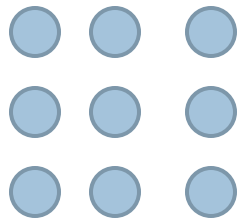
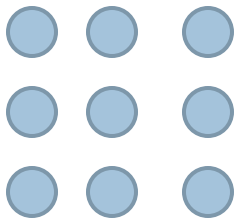
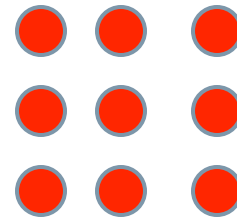
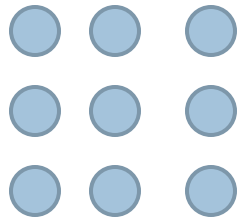
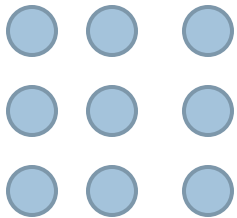
- Train on all groups but one, test on last group
- For each possible combination



Cross-validation

14

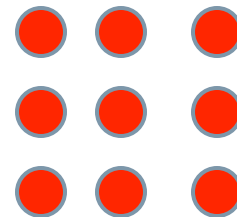
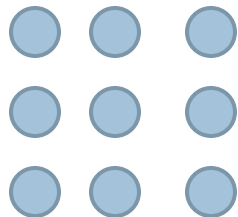
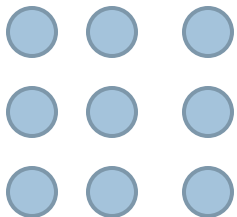
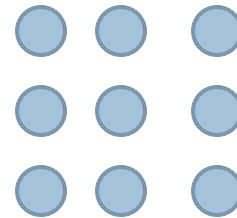
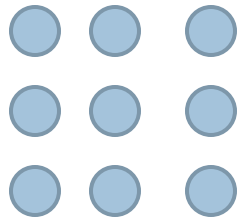
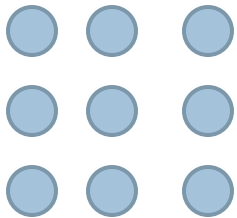
- Train on all groups but one, test on last group
- For each possible combination



Cross-validation

15

- Train on all groups but one, test on last group
- For each possible combination



How many groups?

- K-fold

- ▣ Pick a number K , split into this number of groups

- Leave-out-one

- ▣ Every data point is a fold

How many groups?

- K-fold
 - ▣ Pick a number K , split into this number of groups
 - ▣ Quicker; preferred by some theoreticians

- Leave-out-one
 - ▣ Every data point is a fold
 - ▣ More stable
 - ▣ Avoids issue of how to select folds (stratification issues)

Cross-validation variants

- Flat Cross-Validation
 - ▣ Each point has equal chance of being placed into each fold
- Stratified Cross-Validation
 - ▣ Biases fold selection so that some variable is equally represented in each fold
 - ▣ The variable you're trying to predict
 - ▣ Or some variable that is thought to be an important context

Student-level cross-validation

- Folds are selected so that no student's data is represented in two folds
- Allows you to test model generalizability *to new students*
- As opposed to testing model generalizability *to new data from the same students*

Student-level cross-validation

- Usually seen as the minimum cross-validation needed, in the EDM conference
- Papers that don't pay attention to this issue are usually rejected
 - ▣ OK to explicitly choose something else and discuss that choice
 - ▣ Not OK to just ignore the issue and do what's easiest

Student-level cross-validation

- Easy to do with Batch X-Validation in RapidMiner

Other Levels Sometimes Used for Cross-Validation

- Lesson/Content
- School
- Demographic (Urban/Rural/Suburban, Race, Gender)
- Software Package

Important Consideration

- Where do you want to be able to use your model?
 - New students?
 - New schools?
 - New populations?
 - New software content?

- Make sure to cross-validate at that level

Next Lecture

- More on Generalization and Validity