Linear models

Comparing train and test errors

see video/slides online: https://inria.github.io/scikit-learn-mooc/overfit/learning_validation_curves_slides.html

Video notes

linear = linear in *parameters*, not in predictors. $\Leftrightarrow$ linear combination of inputs

Example: estimating housing prices

- target: sale price
- predictors: living area, year built, N full baths
- linear approximation: explain how the paramaters are interpreted. the parameters are found automatically with the `fit` method in scikit-learn
- slope chosen = minimize distance between prediction and the datapoints (red lines)
- which is itself the squared error -- can compute ourselves with numpy
- then bring this to higher dimensions

For classification: logistic regression

- binary output 0/1
- the model learns to predict a *probability* of one class vs the other
- in the area to the very right or left, the model is very confident about one class vs the other. in between, the model makes less certain predictions.
- strictly speaking, the logistic regression itself is not a linear model, but the log relative probability of one vs the other class is linear in the parameters
- in 2 dimensions: x1, x2 are input variables, colors is the outcome variable
- the straight lines are equi-probability lines in x1-x2 space (see right-hand image: "horizontal cuts through the shape"
- the line is straight because of the linear assumption (linear combination of inputs)

Generalize: multiclass classification

- model will predict probability for each of the three groups (groups are mutually exclusive)

Linear models are not suited to all data

- it's ok as long as data are approximately linearly separable -- example of inductive bias (?). in this case, the model is underfitting
- we could try to solve it with a new feature, or with a different model

Take-home messages

- linear models are simple and fast baselines for regression and classification
- they can underfit when number of features is way smaller than number of samples. in this case, engineering new features can help
- when we have many features, simple linear models are hard to beat.

In [ ]:

In [ ]: