Mathematical Tools for Neuroscience (NEU 314)
Princeton University, Spring 2016
Jonathan Pillow

Lecture 20: Overfitting and Cross-validation

Quick review of likelihood vs. conditional probability:

$P(x|\theta)$ is

- a *conditional probability* when considered as a function of x, the random variable or sample, with θ fixed. (This sums to 1).
- a *likelihood* when considered as a function of θ, the parameter vector, with sample x fixed. (Doesn't sum to 1).

Cross-validation and overfitting

- general goal: generalization performance. We want to extract features of a dataset that will be useful for predicting new data.
- overfitting - phenomenon in which making a model more complex (or adding more parameters) will lead to fitting noise fluctuations in the training data that is *not* useful for predicting new data.
- cross-validation: involves dividing data into a *training set* and a *test set*. Fit the model parameters on the training set and evaluate performance on the test set.
- Training error curve: as we make a model more complex (or add parameters), the training error decreases (for nested models at least: that is, where the more complex model includes the simpler model as a special case, such as adding regressors to a linear regression model.).
- Test error curve: lies above the training error curve (since the parameters are not directly optimized to the test data). As model becomes more complex, exhibits "V" shape: decreases up to a point and then begins increasing again. Region to the left: "underfitting". Region to the right: "overfitting." Sweet spot in terms of model complexity is the point where test error achieves its minimum.
- Also useful for setting regularization parameters governing a model with fixed number of parameters. (e.g., "ridge regression").
- *k-fold cross-validation*: divide data into k blocks. For each of these blocks, train on the other $k-1$ blocks and compute prediction error on the remaining block. Repeat for each of the k blocks as test set, and average the resulting test error curves.
- *Leave-one-out (LOO) cross-validation*: train on all datapoints except for a single held out "test point", and repeat for every datapoint. This is the extreme version of k-fold cross-validation where $k = \#$ of datapoints.

Bayesian vs. Frequentist error bars

- Frequentist: look at the distribution of estimates that would arise if repeated experiments were run from the same (true) point in parameter space. *Confidence intervals*: show range where estimate will land (eg) 95% of the time if repeating the experiment.
- Bayesian: look instead at the posterior distribution over parameters given a single dataset. Requires a prior over parameter space for the posterior to be well-defined. Formal name: *Credible intervals*: refers to interval containing (eg) 95% of the posterior probability mass. (Defined by quantiles of the posterior).
- Next up: bootstrap!