1 Opening

- This year limited to Harvard affiliates only
- stochastic methods and bayesian inference
  - stochastic gradient descent
- stochastic gradient descent

2 Learning Models

- why model
  - prediction
  - explaining
- hypothesis space: potential functional forms
- overfitting: often what's wrong
- generating process is not a straight line.
- OLS was done
- Empirical risk minimization on the points that we have in the sample.
- Sample has to be representative of the population. random sampling.
- error on the entire population (out-of-sample error) is roughly the same as the in-sample error.
- statistics: unbiasedness.
- ML: robustness preferred over unbiasedness.
- neural networks are fit on very large data. regularization is typically used. be skeptical of complex models.
- noisy data. complex model performs worse.
- sample from the population. sampling distribution.
- complex model becomes unstable particularly in the part where data is sparse.
- high variance situation. high variance across random samples.
- K-fold CV. 30 data points. 5-fold 24 data points (more overfitting). over-estimate overfitting than overfitting using 30-data-point sample. more bias (worse performance) with respect to prediction performance
- iterator does not generate data. saves space
- sklearn only three things.
  - transforming. feature engineering. need to create polynomials etc
  - fitting
- sklearn expects list of lists.
- reshape
- columns features. rows observations.
- loading modeling function
- loading assessment function

```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
```

- create an estimator, fit on training data. state changes to a fitted model.
- training set error keeps decreasing with model complexity
- anreas pycon workshop for scikit learn
- training error goes down
- test error goes up and down. artifacts going up down.
- test set has been used at the point we chose.
- error associated with the hyperparameter (degree of polynomials)
- test set was used to choose d. not a good representation of out of sample error.
- test set could be bad.
- we have used up all the data for

3 Validating Models

- validating
- regularization
- hyperparameters two in elastic net
- idiosyncratic validation set is solved with CV

3.1 CV

- deterministic split bad if data is sorted.
- split has to be done up front.
- choose d based on cross validation.
- fit on the entire dataset at that d
- predict and report on held-out part.
- high variance with
- plot the errors individually. not just the mean.
- go back and refit the entire training test set.
- (train-validate)-test set

4 Regularization

- model space restriction
- alpha is prior
- GridSearchCV
- sklearn.dataset
- score: R^2 for regression or accuracy for classification

5 Classification

- regression can be used
- K nearest neighbors
- what to do with probabilities
- scoring function in sklearn splits at 0.5
- asymmetry in scoring (classifying cancer based on probability)