Model Selection:
Training, Test and Validation sets
Cross Validation

Vibhav Gogate
What do we really want?

• Given: A Dataset
• Machine learning: 100 methods
• Why not choose the method with the best fit to the data?
  – Not a good idea because Generalization is important!!
  – It matters how well you classify future unseen data
Example

Which model will I select?
Training-Validation-Test method

• Randomly split the data into
  – Training
  – Validation
  – Test

• Train on training, tune on validation and find how well the tuned model performs on the test data

• Model giving highest accuracy on test wins
Cross Validation

Recycle the data!
**LOOCV** (Leave-one-out Cross Validation)

Let say we have $N$ data points.

$k$ be the index for data points $k=1..N$

Let $(x_k, y_k)$ be the $k^{th}$ record.

Temporarily remove $(x_k, y_k)$ from the dataset.

Train on the remaining $N-1$ Datapoints.

Test your error on $(x_k, y_k)$.

Do this for each $k=1..N$ and report the mean error.
There are N data points. Do this N times. Notice the test data is changing each time.
LOOCV (Leave-one-out Cross Validation)

There are N data points.
Do this N times. Notice the test data is changing each time.
K-fold cross validation

In 3 fold cross validation, there are 3 runs.
In 5 fold cross validation, there are 5 runs.
In 10 fold cross validation, there are 10 runs.

the error is averaged over all runs