# Class 20

October 27, 2022

## Contents



## Overfitting and Generalization

- Model becomes too specific to training set
  - performance on an independent test set becomes poor
    - **poor generalization**

Even Linear Models can overfit

- because less data
- or too much dimensionality in data

### Solutions to prevent overfitting

- Use of simple classifiers: Linear Models
- Large-Margin Classification, SVMs
- Ensemble Classifiers
- Dimensionality Reduction

---

Some minimally expected fundamental rules in ML:

1. Collect Data
   - Avoiding biases while collecting data is extremely important.
2. Data Cleanup
   - Avoiding biases here is essential as well.
   - $n$ samples of labelled data.
3. Training-Validation-Testing
   - We split the labelled data into

    - $t$ samples of Training Data,
    - $v$ samples of Validation Data, and
    - $n - t - v$ samples of Test Data.
  - How you do this split is also important.
  - We run a loop of training on the Training Data, testing on Validation Data, Tuning the hyperparameters then Train again.
  - After some iterations, we might stop and test on the Test Data.
  - If the distribution of Training and Test Data are different then the model will perform very poorly.
    - Split should not change distribution of test data (from training data) but should also be convenient and sufficiently random.
    - Labelled data split *randomly* to training and test data. Then we keep some part of the training data as validation data (random split again).

**Removing the potential bias**

- Cross validation
  - Repeat training and validation multiple number of times. Each time taking a different portion of training set as validation set.
- Model Accuracy
  - Can be given as an error rate
  - Find errors over the test set
  - Get mean accuracy and S.D. accuracy
  - Then accuracy is like $\mu \pm \sigma$ %.

#### k-fold cross validation

The training set will be split $k$ times. Good value for $k$ is 4 or 5. But make sure to not use 10 or 11 (if your data is small)

## Evaluating Classification

**Binary Classification**

| $N = 150$ | Predicted -ve | Predicted +ve | Total |
|---|---|---|---|
| Actual -ve | TN $= 61$ | FP $= 5$ | 66 |
| Actual +ve | FN $= 7$ | TP $= 77$ | 84 |
| Total | 68 | 82 | |

- Accuracy $= (61 + 77)/150$
- Misclassification $= (7 + 5)/150$
- TP rate or hit $= 77/84$
- FP rate or false alarm $= 5/66$
- FN rate or miss $= 7/84$
- TN rate or genuine reject $= 61/66$
- Precision $= 77/82$
- Recall (TPR) $= 77/84$