CS 1675 Introduction to ML

Lecture 4

Evaluation of ML algorithms

Milos Hauskrecht
milos@cs.pitt.edu
5329 Sennott Square, x4-8845

people.cs.pitt.edu/~milos/courses/cs1675/

Homework assignments

- **Homework assignment 1 due today**
- **Homework assignment 2 is out and due on Thursday, January 26, 2017**

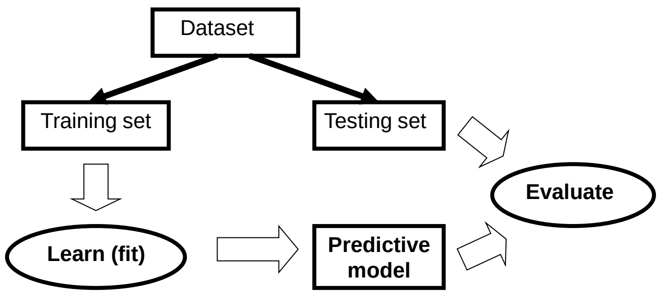
Two parts: **Report + Programs**

Learning process (second look)

1. Data
   - Understand the source of data
   - Real data may need a lot of cleaning/preprocessing
2. Model selection:
   - How to pick the models: manual/automatic methods
3. Choice of the objective (error or loss) function
   - Many functions possible: Squared error, negative log-likelihood, hinge loss
4. Learning:
   - Find the set of parameters optimizing the error function
5. Application/Testing:
   - Evaluate on the test data
   - Apply the learned model to new data

Evaluation of models

- Simple holdout method
  - Divide the data available to the training and test data



  - Typically 2/3 training and 1/3 testing

Evaluation measures

Regression model f: X → Y where Y is real valued

- Mean Squared Error

$$MSE(D, f) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$$

- Mean Absolute Error

$$MAE(D, f) = \frac{1}{n}\sum_{i=1}^{n}|y_i - f(x_i)|$$

- Mean Absolute Percentage Error

$$MAPE(D, f) = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - f(x_i)}{y_i}\right|$$

Evaluation measures

Regression model f: X → Y where Y is real valued

- The error is calculated on the data D, say

$$MSE(D, f) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$$

- This quantity is an estimate of the true error for *f*

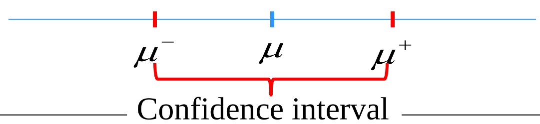
Important question:

- How close is our estimate to the true mean error?

To answer the question we need to resort to statistics:

- How confident we are the true error falls into interval around our estimate μ ?

Answer: with probability 0.95 the true error is in interval $[\mu^-, \mu^+]$

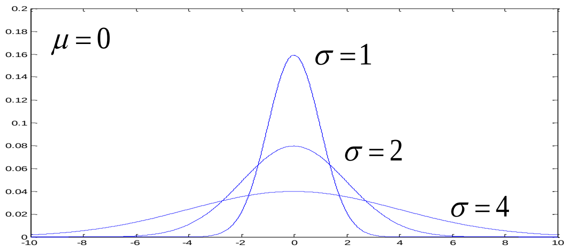


Evaluation

- **Central limit theorem:**

  Let random variables $X_1, X_2, \cdots X_n$ form a random sample from a distribution with mean μ and variance σ , then if the sample n is large, the distribution

$$\sum_{i=1}^{n} X_i \approx N(n\mu, n\sigma^2) \quad \text{or} \quad \frac{1}{n}\sum_{i=1}^{n} X_i \approx N(\mu, \sigma^2 / n)$$



Statistical significance test

- **Statistical tests for the mean**
  - **H0 (null hypothesis)** $\mu^0 = \mu^*$
  - **H1 (alternative hypothesis)** $\mu^0 \neq \mu^*$
- **Basic idea:**

  we use the sample mean and check how probable it is to occur given that the true mean is 0

$$E[X] = \mu^*$$

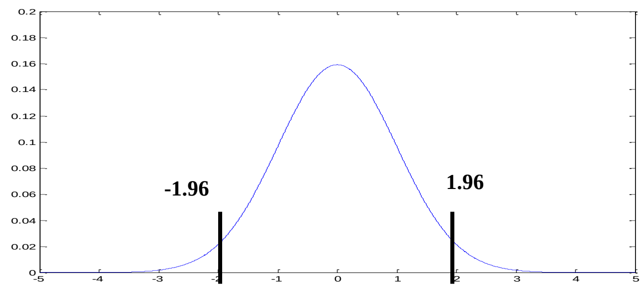
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

If the probability that $\overline{X}$ comes from the normal distribution with mean μ* is small – we reject the null hypothesis on that probability level

Statistical significance test

- **Statistical tests for the mean**
  - **H0 (null hypothesis)** $\mu^0 = \mu^*$
  - **H1 (alternative hypothesis)** $\mu^0 \neq \mu^*$
- **Assume we know standard deviation** σ

$$z = \frac{\bar{X} - \mu^*}{\sigma}\sqrt{n} \approx N(0,1) \quad \text{with} \quad \text{P=0.95} \quad z \in [-1.96, 1.96]$$

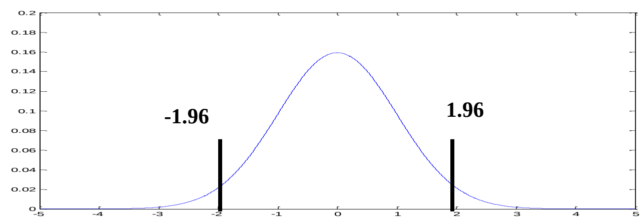


Statistical significance test

- **Statistical tests for the mean**
  - **H0 (null hypothesis)** $\mu^0 = \mu^*$
- **Assume we know standard deviation** σ

$$z = \frac{\bar{X} - \mu^*}{\sigma}\sqrt{n} \approx N(0,1) \quad \text{with} \quad \text{P=0.95} \quad z \in [-1.96, 1.96]$$

- **Z-test: If z is outside of the interval – reject the null hypothesis at significance level 5 %**



Statistical significance test

- **Statistical tests for the mean**
  - **H0 (null hypothesis)** $\mu^0 = \mu^*$
- **Problem: we do not know the standard deviation** σ
- **Solution:**

$$t = \frac{\overline{X} - \mu^*}{s}\sqrt{n} \approx \text{t} - \text{distribution} \quad \text{(Student distribution)}$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

- Estimate of the standard deviation

- **T-test: If t is outside of the tabulated interval reject the null hypothesis at the corresponding significance level**

Confidence in the estimate

The statistical significance test lets us answer:

- The probability with which the true error falls into the interval around our estimate, say :

$$MSE(D, f) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$$

- Compare two models M1 and M2 and determine based on the error on the data entries the probability with which model M1 is different (or better) than M2

$$MSE(D, f_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_1(x_i))^2 \qquad MSE(D, f_2) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_2(x_i))^2$$

Trick:

$$MSE(D, f_1) - MSE(D, f_2) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_1(x_i))^2 - \frac{1}{n}\sum_{i=1}^{n}(y_i - f_2(x_i))^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i - f_1(x_i))^2 - (y_i - f_2(x_i))^2$$

Evaluation measures

Similarly evaluations measures and can be defined for the classification tasks

Assume binary classification:

		Actual	
		Case	**Control**
Prediction	**Case**	**TP** 0.3	**FP** 0.1
	Control	**FN** 0.2	**TN** 0.4

Misclassification error:

$$E = FP + FN$$

Misclassification error:

$$E = FP + FN$$

Sensitivity:

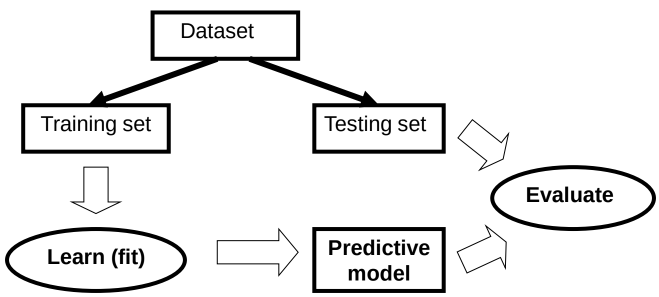
$$SN = \frac{TP}{TP + FN}$$

Specificity:

$$SP = \frac{TN}{TN + FP}$$

Evaluation of models

- **We started with a simple holdout method**



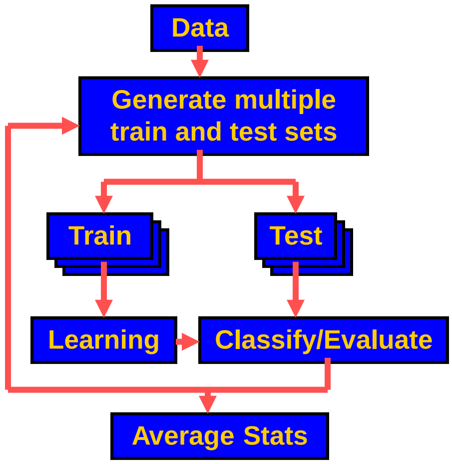
Problem: the mean error results may be influenced by a lucky or an unlucky **training and testing** split especially for a small size D

Solution: try multiple train-test splits on D and average their results

Evaluation of models via random resampling

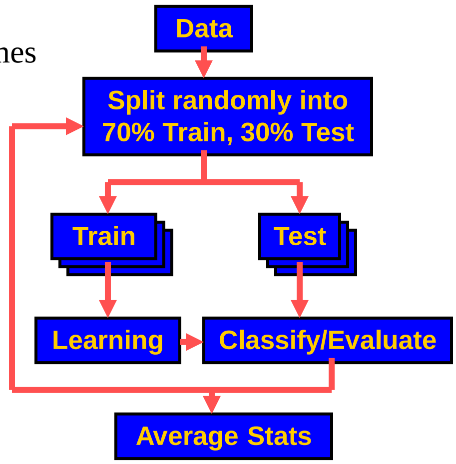
Other more complex methods

- Use multiple train/test sets
- Based on various random re-sampling schemes:
  - **Random sub-sampling**
  - **Cross-validation**
  - **Bootstrap**



Evaluation of models using random subsampling

- **Random sub-sampling**
  - Repeat a simple holdout method k times

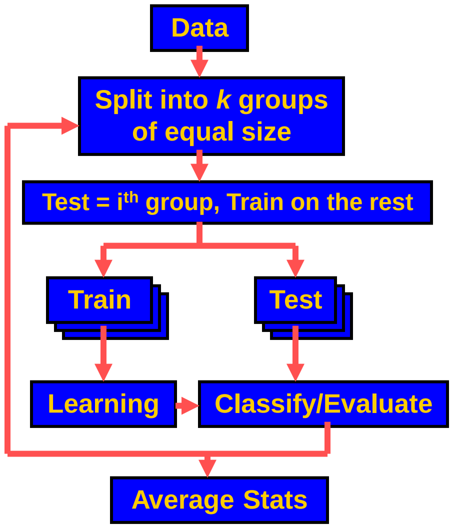


Evaluation of models using k-fold cross-validation

Cross-validation (k-fold)

- Divide data into k disjoint groups, test on k-th group/train on the rest
- Typically 10-fold cross-validation
- Leave one out cross-validation

  (k = size of the data D)



Evaluation of models using bootstrap

Bootstrap

- The training set of size N = size of the data D
- Sampling with the replacement