**White Board Notes for Lectures 1 & 2**

**Lecture 1:**

* **if independent**
  + ***H:* Hypothesis**
  + ***E:* Evidence**
    - **Example: H = fever, E = flu**
    - **Is P(flu|fever) easier to predict or P(fever|flu)**
    - **P(fever|flu) is easier to calculate**
* **Naïve Bayes**
  + **Usually a good baseline model**
  + **The ‘Naïve” assumption: assumption of independency** 
    - **simplifies Baye’s rule because it deals with unseen instances with a specific evidence**
    - **also deals with curse of dimensionality**
* **Entropy**
  + - **If P(Y) or P(N) = 1 🡪 E = 0 (no noise, but pure)**

**Lecture 2: Performance Evaluations**

1. **Error\_rate =**
2. **Sampling Techniques**
   1. **Training: resubstitution error**
   2. **Testing: Generalization**
   3. **Validation: Tune Parameters**
      1. **Note: Testing/Validation terminology can be used interchangeably, as long as it is specifically specified which is for what**
   4. **Which algorithms needs to be tuned?**
      1. **Ex. Decision Trees**
         1. **np = # of cases per parent**

**Ex: np = 20:**

25

19

21

4

* + - 1. **nc = # of cases per child**

25

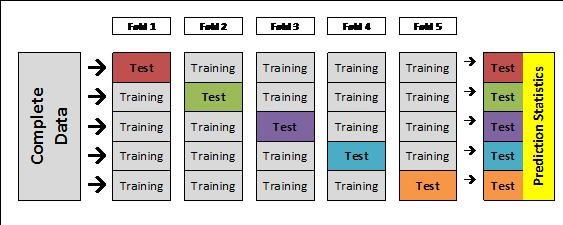
**Ex: nc = 6**

4

21

* + 1. **Other Ex. K-NN, SVM, Neural Nets**
  1. **Hold Out**
     1. **Typically 66% training / 34% testing for large datasets**
     2. **Good for balanced dataset**
  2. **Stratified Holdout**
     1. **For large and unbalanced datasets**
        1. **C1: malignant 100 cases 🡪 P(C1)**
        2. **C2: benign 1000 cases 🡪 P(C2)**
  3. **Repeated Holdout (About 30 times)**
     1. **If dataset is small & balanced**
     2. **Can cause overlapping in test sets**
  4. **Cross Validation**
     1. **Useful for small datasets**
     2. **solves the issue of overlapping test sets from repeated holdout**

**Ex: 5-Fold CV**



* 1. **Stratified Cross Validation**
     1. **If dataset is small and is unbalanced**
  2. **Leave-One-Out**
     1. **Used on even smaller datasets**
     2. **train on n-1 & test on the observation that was left out**
     3. **Terrible on an unbalanced dataset**
     4. **Cannot use stratification with Leave-One-Out technique**
  3. **Bootstrapping**
     1. **Very small and unbalanced datasets**
     2. **Samples with Replacement**
     3. **: is any performance measuer e.g. Accuracy, Sensitivity, etc**
     4. **a.k.a .632 Bootstrap**
        1. **samples**
     5. **needs to be repeated for more reliability (performance is averaged)**

1. **Statistical Reliability**
   1. **If Algorithm 1 produced 74% accuracy, Algorithm 2 produced 72% accuracy, which one is more reliable? How close are these to the true error rate?**
   2. **Based off of Bernoulli Process**

|  |  |  |
| --- | --- | --- |
|  | **Actual** | **Predicted** |
| **Correctly classified** | **C1** | **C1** |
| **Incorrectly classified** | **C1** | **C2** |

1. **🡨 more samples, less variance CLT**
   1. **Confidence Intervals**
      1. ***X = age***

|  |
| --- |
| **20** |
| **31** |
| **18** |
| **32** |
| **39** |

**Want sample mean of X to be close to the true mean**

* + 1. **Can give an confidence interview of how confidence your sample mean is within the true mean**
    2. **where**

1. **Paired T-test**
   1. **Used to compare algorithms e.g. Performance of Decision Trees vs Performance of Naïve Bayes**
   2. **Hypothesis Testing (Example)**

* + 1. **Calculate Test statistic**
       1. **T-test for small data set**
       2. **Z-test for larger data set**
    2. **P-Value**
    3. **Accept or Reject Null Hypothesis**
       1. **If**
       2. **If**
       3. **is considered “Highly Significant”**

1. **Loss Functions**
   1. **Success-rate a.k.a. 0-1 Loss function**
   2. **Quadratic Loss Function**
   3. **Information Loss Function**

**🡨 Therefore need the negative in front**