CS 2750 Machine Learning

Lecture 4

Evaluation of predictors and learners

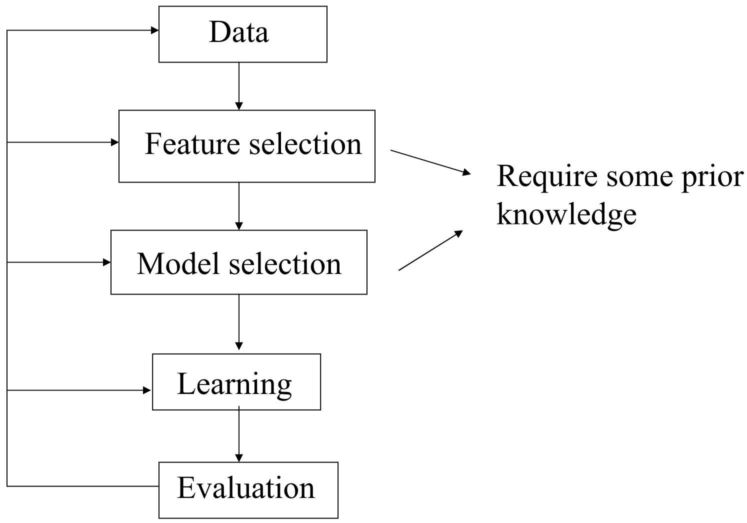
Milos Hauskrecht
milos@cs.pitt.edu
5329 Sennott Square, x4-8845

http://www.cs.pitt.edu/~milos/courses/cs2750/

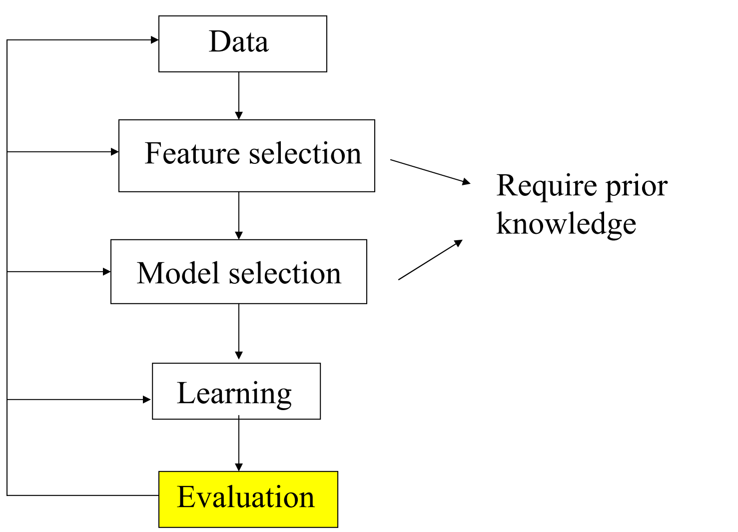
Administration

- **Homework 1. due today**
- **Homework 2 is out. Due next week on Wednesday.**

Design cycle



Design cycle



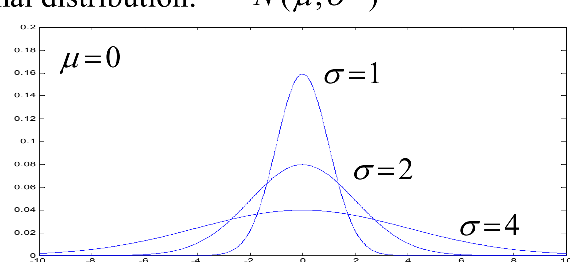
Evaluation.

- **Evaluation:**
  - **Use pristine test data held out from the data set.**
    - **Reason:** Overfit can cause the training error to go to zero so it makes sense to evaluate only on the test error.
  - **More complex alternative: cross-validation**
- **Three evaluation questions:**
  - **Question 1:** How far is the test error from the true error?
    - test error approximates the generalization (true) error
  - **Question 2.** How do we compare two different classifiers? Which one is better than the other?
  - **Question 3.** How do we compare two different learning algorithms? Which one is better than the other?

How far is the test error from the true error?

- **Problem:** we cannot be 100 % sure about the goodness of the test error approximation
- **Solution:** statistical methods, confidence intervals
- It is based on:
  - **Central limit theorem:** the sum of a large number of random samples is normally distributed.

Normal distribution: $N(\mu, \sigma^2)$

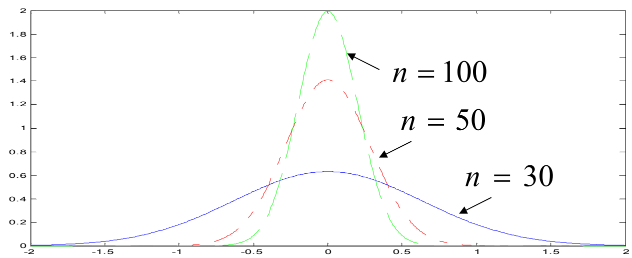


Central limit theorem

- **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^2, then if the sample n is large, the distribution

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

Effect of increasing the sample size *n* on the sample mean:



$\mu = 0$

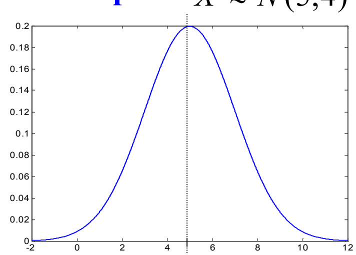
$\sigma^2 = 4$

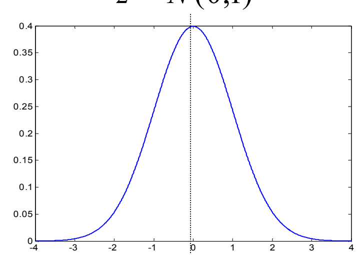
Transformation to N(0,1)

- **Sample mean:** $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \approx N(\mu, \sigma^2 / n)$
  - Is normally distributed around the true mean
- **We can transform the sample mean as follows:**

$$z = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \approx N(0,1)$$

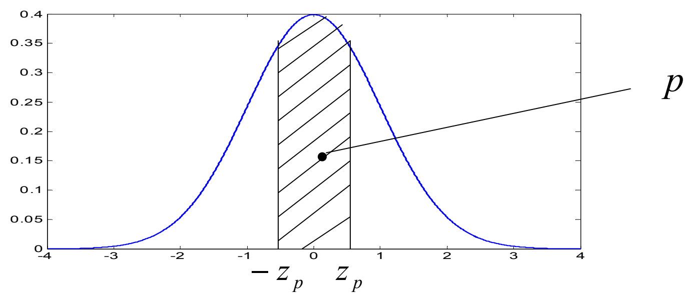
- **Example:** $\bar{X} \approx N(5,4)$ $\qquad z = N(0,1)$





Confidence intervals

- **Assume N(0,1)**
- **We are interested in:**
  - Finding the symmetric interval **around the mean** such that the probability of seeing a sample from it is *p*
  - Measuring the distance of end points from 0 in terms of $\sigma = 1$

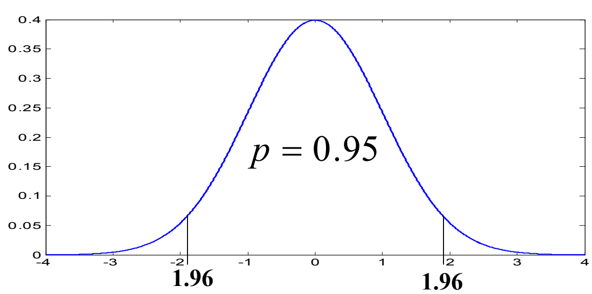


$p \rightarrow [-z_p, z_p]$

Confidence intervals

- **Assume N(0,1):** $p \rightarrow [-z_p, z_p]$
- **Values** (p, z_p) **are tabulated**

- **Example:** $p = 0.95 \rightarrow z_p = 1.96$



- **With confidence 0.95 we see values in interval** $[-1.96, 1.96]$

Confidence intervals

- **Back to case:** $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \approx N(\mu, \sigma^2 / n)$
- Probability mass under the normal curve for a symmetric interval around the mean is invariant when interval distances are measured in terms of the standard deviation
- **For** $N(0,1)$ $\quad p = 0.95 \Longrightarrow z_p = 1.96$
- **For** $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \approx N(\mu, \sigma^2 / n)$

$$z = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \approx N(0,1) \qquad \bar{X} \in [\mu - z_p \frac{\sigma}{\sqrt{n}}, \mu + z_p \frac{\sigma}{\sqrt{n}}]$$

$$p = 0.95 \Longrightarrow \bar{X} \in [\mu - 1.96(\sigma / \sqrt{n}), \mu + 1.96(\sigma / \sqrt{n})]$$

$$\Longrightarrow \mu \in [\bar{X} - 1.96(\sigma / \sqrt{n}), \bar{X} + 1.96(\sigma / \sqrt{n})]$$

Confidence interval

- **Problem:** But typically the variance is not known
- **Solution:** estimate the variance from the sample

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

- Assume the sample mean falls into the interval centered at the mean:

$$\bar{X} \in \left[\mu - t_p \frac{s_n}{\sqrt{n}}, \mu + t_p \frac{s_n}{\sqrt{n}}\right]$$

- Or equivalently that the mean falls into the interval centered around the sample mean:

$$\mu \in \left[\bar{X} - t_p \frac{s_n}{\sqrt{n}}, \bar{X} + t_p \frac{s_n}{\sqrt{n}}\right]$$

- **This happens with some probability *p* that depends on** t_p

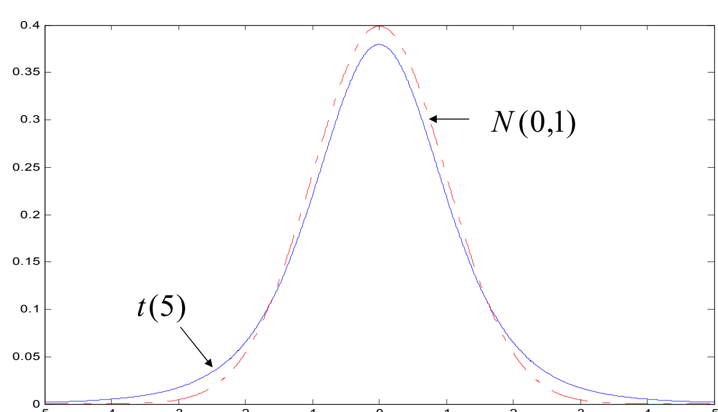
Confidence interval

- **Let:** $t = \frac{\bar{X} - \mu}{s_n}\sqrt{n}$
- The difference from the known variance case:
  - t is not normally distributed, instead it follows a **Student distribution** (t distribution)
  - Student distribution has one additional parameter: the **degree of freedom**
  - **For** $s_n = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$ **t has n-1 degrees of freedom**

$$t(n-1) = \frac{\bar{X} - \mu}{s_n}\sqrt{n} \approx \text{t distribution (n - 1)}$$

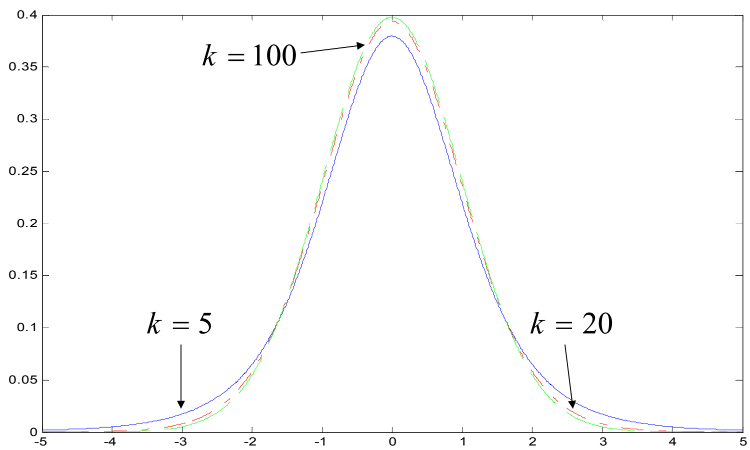
Student distribution

- Student distribution versus normal N(0,1)



Student distribution

- Student distribution with k degrees of freedom
  - For $k \to \infty$ it approaches N(0,1)



So how different the test error can be?

- **Select confidence level (probability) (e.g. p=0.95)**
- **Compute interval into which the sample mean falls** with that confidence:
  - **For unknown mean and know variance**

$$\bar{X} \in [\mu - z_p \frac{\sigma}{\sqrt{n}}, \mu + z_p \frac{\sigma}{\sqrt{n}}] \quad \textbf{and} \quad \mu \in [\bar{X} - z_p \frac{\sigma}{\sqrt{n}}, \bar{X} + z_p \frac{\sigma}{\sqrt{n}}]$$

E.g. for p=0.95 $\mu \in [\bar{X} - 1.96(\sigma / \sqrt{n}), \bar{X} + 1.96(\sigma / \sqrt{n})]$

  - **For unknown mean and unknown variance**

$$\bar{X} \in [\mu - t_p(n-1)\frac{s_n}{\sqrt{n}}, \mu + t_p(n-1)\frac{s_n}{\sqrt{n}}] \quad \textbf{and}$$

$$\mu \in [\bar{X} - t_p(n-1)\frac{s_n}{\sqrt{n}}, \bar{X} + t_p(n-1)\frac{s_n}{\sqrt{n}}]$$

  - **E.g. for p=0.95 and n=30**

$$\mu \in [\bar{X} - 2.045 \frac{s_n}{\sqrt{n}}, \bar{X} + 2.045 \frac{s_n}{\sqrt{n}}]$$

Comparison of two predictors

Predictor 1 uses function $f_1(\mathbf{x})$ to predict ys

Predictor 2 uses function $f_2(\mathbf{x})$ to predict ys

- Test data are used to approximate the **true errors**

$$Error_1 = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_1(\mathbf{x}_i))^2$$

$$Error_2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_2(\mathbf{x}_i))^2$$

Test errors

- **Assume that:** the sample size n is sufficiently large
- **Assume that we observed :** $Error_1^0 > Error_2^0$

  or that $\Delta E^0 = Error_1^0 - Error_2^0 > 0$
- **Question:** How sure are we that the predictor 2 is better than the predictor 1 in terms of true errors ?

Comparison of two predictors

- **True errors:**

$$Error_1^{True} = E_{(\mathbf{x},y)}\left[(y - f_1(\mathbf{x}))^2\right]$$

$$Error_2^{True} = E_{(\mathbf{x},y)}\left[(y - f_2(\mathbf{x}))^2\right]$$

- **Predictor 2 is better than Predictor 1 if:** $Error_1^{True} > Error_2^{True}$
  - or $\mu_{diff} = E_{(\mathbf{x},y)}\left[(y - f_1(\mathbf{x}))^2 - (y - f_2(\mathbf{x}))^2\right] > 0$
- **Problem:** we do not know the true mean error difference
- **But we can** approximate the last quantity with the sample

$$\Delta E = Error_1 - Error_2$$

$$\Delta Error = \frac{1}{n}\sum_{i=1}^{n}\left[(y_i - f_1(\mathbf{x}_i))^2 - (y_i - f_2(\mathbf{x}_i))^2\right]$$

Paired squared differences for test sample

Comparison of two predictors

True error differences

$$\mu_{diff} = E_{(\mathbf{x},y)}\left[(y - f_1(\mathbf{x}))^2 - (y - f_2(\mathbf{x}))^2\right]$$

Error differences based on the sample of size n

$$\Delta E = \frac{1}{n}\sum_{i=1}^{n}\left[(y_i - f_1(\mathbf{x}_i))^2 - (y_i - f_2(\mathbf{x}_i))^2\right]$$

Assume: X **is a random variable, such that**

$$X_i \approx (y_i - f_1(\mathbf{x}_i))^2 - (y_i - f_2(\mathbf{x}_i))^2$$

But then

$$\Delta E = \overline{X} = \frac{1}{n}\sum_{i=1}^{n}\left[(y_i - f_1(\mathbf{x}_i))^2 - (y_i - f_2(\mathbf{x}_i))^2\right]$$

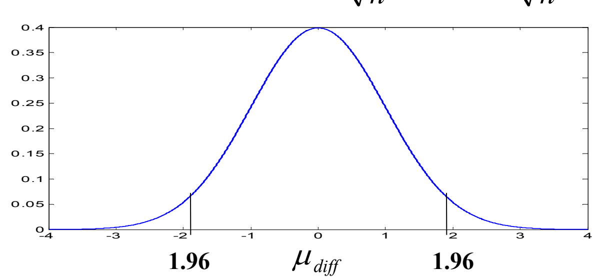
Central limit result testifies about the difference:

$$\Delta E = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \approx N(\mu, \sigma^2 / n)$$ X_i - is a random variable

Comparison of two predictors

- **Assume the variance σ_{diff} is known**
- **Then we can derive a constant z_p such that with a probability *p* our estimate falls into:**

$$\Delta E = \overline{X} \in [\mu_{diff} - z_p \frac{\sigma_{diff}}{\sqrt{n}}, \mu + z_p \frac{\sigma_{diff}}{\sqrt{n}}]$$



- **But we have a different objective here ….**

Comparison of two predictors

- **Our objective is to determine what is the probability that $\mu_{diff} > 0$ holds given an observed $\Delta E^0 > 0$**
- **An alternative formulation: the probability that we can reject $\mu_{diff} \leq 0$ given $\Delta E^0 > 0$**

This is a classic hypothesis testing problem in statistics

- **Typical formulation:**
  - **H0 (null hypothesis)** $\mu_{diff} = 0$
  - **H1 (alternative hypothesis)** $\mu_{diff} \neq 0$
- **Question:** can we reject the null hypothesis with some confidence given the observed sample mean (ΔE^0)of size n
- **The hypothesis** here is undirectional and standard two-sided z-test or t-test can be applied to determine the confidence level for reject

Comparison of two predictors

Our case is different:

- **H0 (null hypothesis)** $\mu_{diff} \leq 0$
- **H1 (alternative hypothesis)** $\mu_{diff} > 0$

- That is, we want to reject the case when the true mean of the score differences is $\mu_{diff} \leq 0$ based on $\Delta E^0 > 0$ with some confidence level.

- This is a directional hypothesis

- **Test methods:**
  - **One-sided z-test** (for the known variance case)
  - **One-sided t-test** (for the unknown variance case)

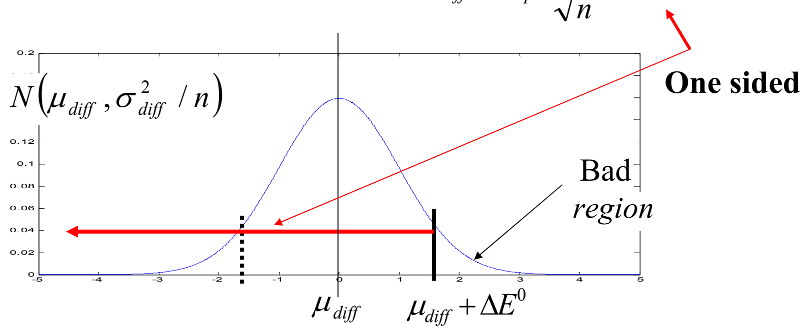
Comparison of two predictors

- **Support for an alternative hypothesis**

$$P(\mu_{diff} > 0) = P(\Delta E < \mu_{diff} + \Delta E^0)$$

- **From the central limit:** $P(\Delta E < \mu_{diff} + z_p^1 \frac{\sigma_{diff}}{\sqrt{n}}) = p^1$

$$\Delta E \approx N\left(\mu_{diff}, \sigma^2_{diff} / n\right)$$



- **Computation:** $\Delta E^0 = z_p^1 \frac{\sigma_{diff}}{\sqrt{n}} \Rightarrow z_p^1 = \Delta E^0 \frac{\sqrt{n}}{\sigma_{diff}} \Rightarrow p^1$

Example

- **Example:** $\Delta Error^0 = 0.1, (\sigma_{diff} / \sqrt{n}) = 0.061$

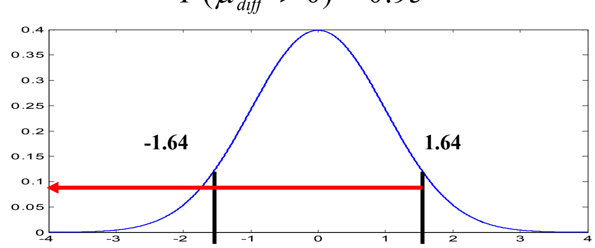
$$P(\mu_{diff} > 0) = ?$$

- **Then:**

$$\Delta Error^0 = z_p^1 \frac{\sigma_{diff}}{\sqrt{n}} \Rightarrow z_p^1 = \Delta Error^0 \frac{\sqrt{n}}{\sigma_{diff}} \approx 1.64$$

- **Distance of 1.64 standard deviations corresponds to one sided *p* value of 0.95**

$$P(\mu_{diff} > 0) = 0.95$$



Comparison of two predictors

- **Case: unknown standard deviation** σ_{diff}
- **Solution:** use the estimate of the standard deviation

$$s_{diff}^{n} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$ - Estimate of the standard deviation

$$t(n-1) = \frac{\overline{X} - \mu_{diff}}{s_{diff}^{n}} \sqrt{n} \approx \text{t distribution}$$

- **Compute the probability of a one sided interval:**

$$P(\overline{X} < \mu_{diff} + t_p^1(n-1)\frac{s_{diff}^{n}}{\sqrt{n}}) = p^1$$

$$\Delta Error^0 = t_p^1(n-1)\frac{s_{diff}^{n}}{\sqrt{n}} \Rightarrow t_p^1(n-1) = \Delta Error^0 \frac{\sqrt{n}}{s_{diff}^{n}} \Rightarrow p^1$$

Comparison of two algorithms

Comparison of two learning algorithms L1 & L2 can be a much harder task, especially when data are small.

- **Problem:** Learning can be performed on many different training sets
  - One training set may not fit well one algorithm, but on average it can perform better.
- **Optimal evaluation settings:**
  - draw a sequence of *k* independent training and testing sets.
  - Evaluate & compare methods based on average of errors for every train-test cycle
- **Practical evaluation settings:**
  - we do not have the luxury of independent samples
  - use surrogate sampling with dependencies**: cross-validation, re-sampling**