Comparing classifiers

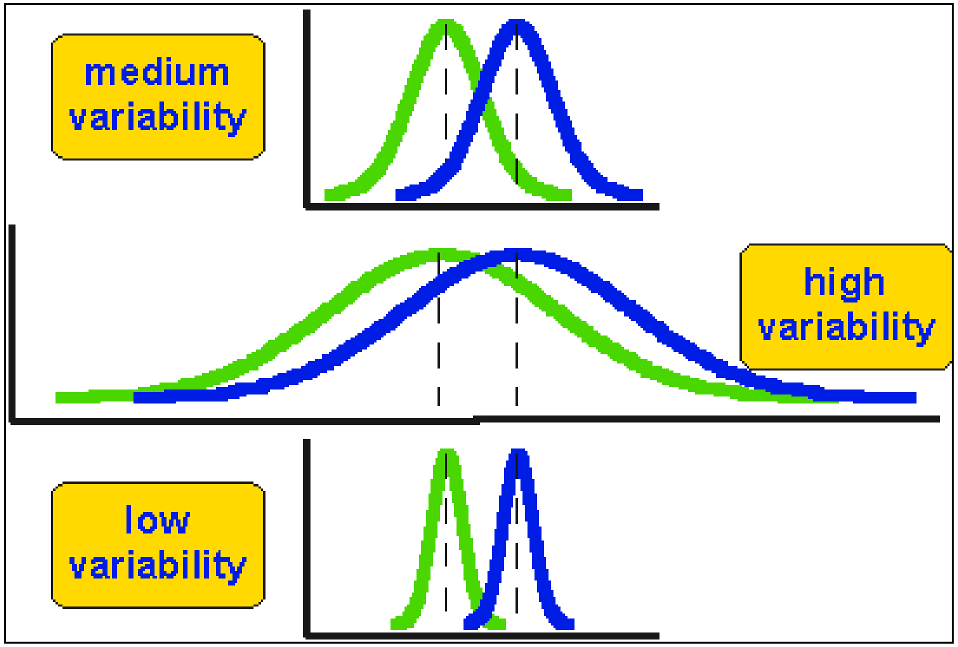
Lecture 9

Outline

- Performance measure: error rate
- Generating test set
- Predicting performance interval
- Comparing two classifiers
- Cost-based evaluation

Comparing data mining schemes

- Which of two learning schemes perform better?
- Note: this is domain dependent!
- Obvious way: compare error (success) rate on different test sets (for example, for different folds of cross-validation)
- Problem: variance in estimate



Statistical test for difference

- Question: whether the means of two samples are *significantly* different.
- In our case the samples are cross-validation accuracy for different folds from the same dataset
- The same Cross Validation is applied twice: once for classifier A and once for classifier B

Probability distribution of sampling means

- Let m_X denote the mean of the probability of success of classifier A, and m_Y – the mean of the probability of success of classifier B

- We already know that the means of multiple samplings for each classifier are normally distributed around the real means μ_A and μ_B of classifier's correctness for the entire population

Probability distribution of sample mean differences

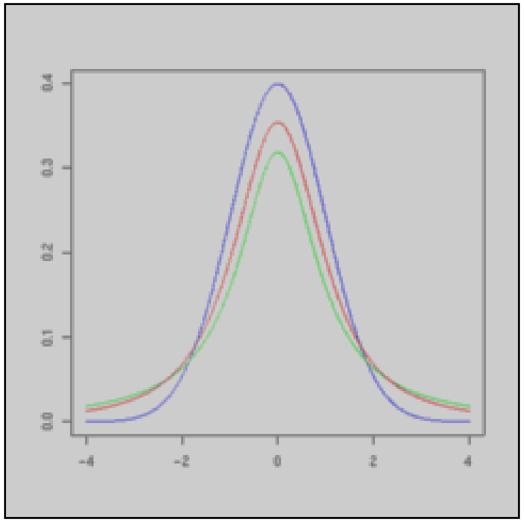
- We could estimate the intervals for the real means μ_A and μ_B for a certain confidence level
- *Suppose,* $\mu_A = 70 \pm 10$ *and* $\mu_B = 60 \pm 10$

- Which one is better?

Real means are somewhere inside these intervals.
Maybe they are just the same?

Probability distribution of sample mean differences

- If we take multiple samplings, and for each sample compute the difference of the means d_m, then for multiple samplings the distribution of the mean differences approaches the *Student's* distribution T with k-2 degrees of freedom



Student's distribution (red) for 2 degrees of freedom compared to normal distribution (blue)

Standard deviation of Student's distribution

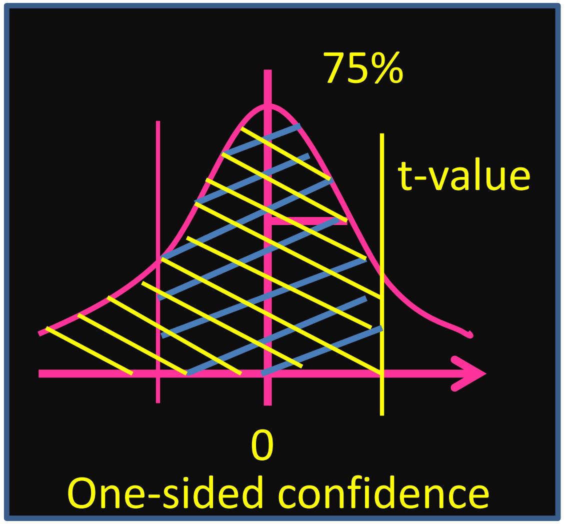
- Student's distribution is very similar to the normal distribution. Not surprisingly, its mean represents a mean of a real difference between X and Y for the entire population, μ_d , and its standard deviation is inversely proportional to the sample size *N*:
- $\sigma_d^2 = s_d^2/N$

Null-hypothesis

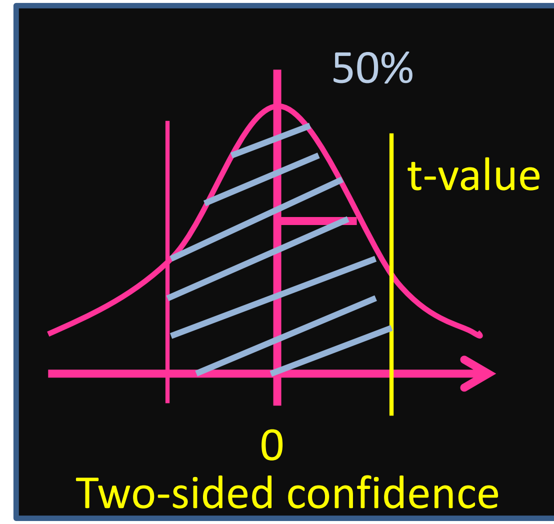
- We formulate our statistical hypothesis about the true value of μ_d:

$$\mu_d=0$$

Next, we select the level of significance (or confidence), and we find how many standard deviations from the mean $\mu_d=0$ should be sample mean difference m_d of any random sampling in order to be still considered 0-difference (no statistically significant difference)



T-table

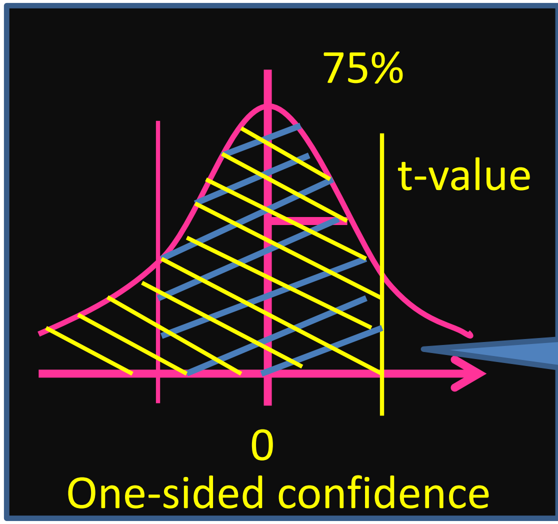


How many standard deviations from the mean – *t*-value

Degrees of freedom

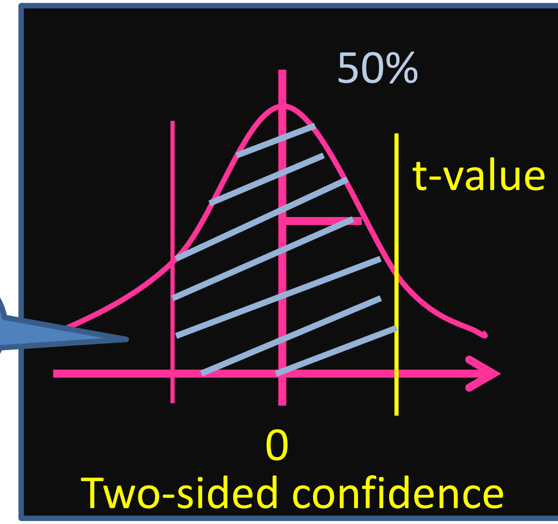
One Sided	**75%**	**80%**	**85%**	**90%**	**95%**	**97.5%**	**99%**	**99.5%**	**99.75%**	**99.9%**	**99.95%**
Two Sided	**50%**	**60%**	**70%**	**80%**	**90%**	**95%**	**98%**	**99%**	**99.5%**	**99.8%**	**99.9%**
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437

T-table



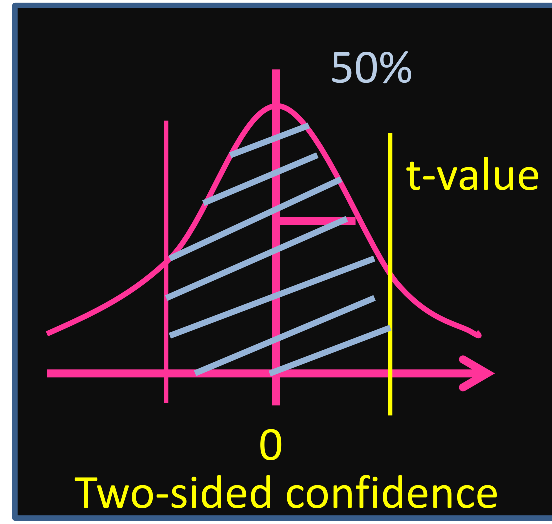
α=25%

α=50%



- One-sided test is used if we only interested if our difference is significantly greater than zero, or significantly smaller than zero, but not both
- Two-sided – if we are interested if our difference is significantly different from zero – both greater and smaller

T-test



- If the mean of differences of two samples is within the interval, then our Null-hypothesis is correct – there is no significant difference between two classifiers (for a given significance level)

- If the mean of differences is outside the interval, then the difference is significant (not by random chance), and we select the classifier with higher on average correctness

Comparing performance of two classifiers in practice

- Perform *k* classifications on each of k datasets using classifier A and classifier B in turn
- Compute difference of classification means for each dataset
- Find mean (average) and variance *s* of differences
- Fix a significance level α. Compute confidence for two-sided T-distribution: C=1.00 – α. Find t-value from the T-table for confidence C and k-2 degrees of freedom
- Find interval for the hypothesis μ_d=0: $\mu_d = 0 \pm t \frac{\sigma}{\sqrt{N}}$
- If the mean of differences is greater than $+t \frac{\sigma}{\sqrt{N}}$, then the first classifier is significantly better,
- if the mean of differences is less than $-t \frac{\sigma}{\sqrt{N}}$, then the second classifier is significantly better

Example. Input

- We have compared two classifiers through cross-validation on 10 different datasets (folds).
- The success rates are:

Dataset	Classifier A	Classifier B	Difference
1	89.4	89.8	-.4
2	90.2	90.6	-.4
3	87.7	88.2	-.5
4	90.3	90.9	-.6
5	91.2	91.7	-.5
6	89.4	89.8	-.4
7	90.2	90.6	-.4
8	87.7	88.3	-.5
9	90.3	90.9	-.6
10	91.2	91.7	-.5

Example. Mean and variance of differences

- m_d = -0.48
- s_d = 0.0789

$$\sigma_d = \frac{s_d}{\sqrt{k}} = \frac{0.0789}{\sqrt{10}} = 0.0249$$

Example. T-interval

σ_D=0.0249

The critical value of *t* for a two-tailed statistical test, α = 10% (c=90%) and k-2=8 degrees of freedom is: 1.86

The average difference should be outside the interval [-1.86*0.0249, 1.86*0.0249] in order to be significant

One Sided	**75%**	**80%**	**85%**	**90%**	**95%**	**97.5%**	**99%**	**99.5%**	**99.75%**	**99.9%**	**99.95%**
Two Sided	**50%**	**60%**	**70%**	**80%**	**90%**	**95%**	**98%**	**99%**	**99.5%**	**99.8%**	**99.9%**
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781

Example. Solution

Significance α = 10%:

The average difference should be outside interval [-0.046, 0.046] in order to be significant

Our average difference is -0.48. The second classifier is significantly better than the first