Evaluation

Rotem Dror
rtmdrr@seas.upenn.edu|http://rtmdrr.github.io

Slides were created by Dan Roth (for CIS519/419 at Penn or CS446 at UIUC), Rotem Dror, or other authors who have made their ML slides available.

Penn Engineering

Administration (9/30/20)

Are we recording? YES!

Available on the web site

- Remember that all the lectures are available on the website before the class
  - Go over it and be prepared
- HW 1 is out
  - Covers: SGD, DT, Feature Extraction, Ensemble Models, & Experimental Machine Learning
  - Start working on it now. Don't wait until the last day (or two) since it could take a lot of your time
- Go to the recitations and office hours
- Questions?
  - Please ask/comment during class.
  - Give us feedback

Have you started to work on HW1?

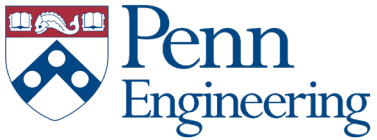
Yes, basically done

Yes; not done yet, but it's going well.

Yes, but I have a lot of questions.

No, but I read it and it seems easy enough. Haven't had a chance to look at it yet...

Metrics
Methodologies
Statistical Significance



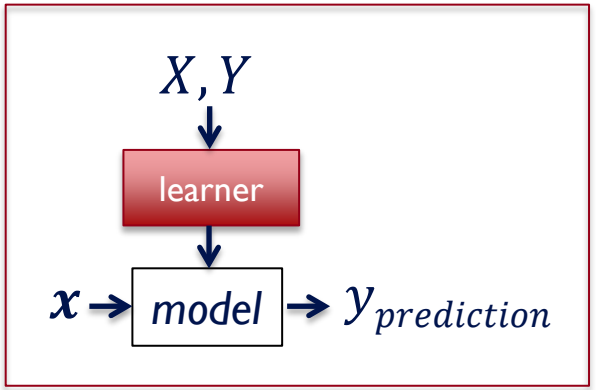
Flow of Batch Machine Learning

Given: labeled training data $X, Y = \{< \boldsymbol{x}_i, y_i >\}_{i=1}^{n}$

- Assumes each $x_i \sim D(X)$ with $y_i = f_{target}(\boldsymbol{x}_i)$

Train the model:

$model \leftarrow classifier.train(X, Y)$



Apply the model to new data:

- Given: new unlabeled instance $\boldsymbol{x} \sim D(X)$

  $y_{prediction} \leftarrow model.predict(\boldsymbol{x})$

Key questions:

How to determine the quality of the model?

(i) measuring performance

(ii) understanding the significance of the results (is it better than other models?)

Metrics

Methodologies

Statistical Significance

Penn Engineering

Metrics

- We train on our training data Train = $\{x_i, y_i\}_{1,m}$
- We test on Test data.
- We often set aside part of the training data as a development set, especially when the algorithms require tuning.
  - In the HW we asked you to present results also on the Training; why?
- When we deal with binary classification we often measure performance simply using Accuracy:

$$\text{accuracy} = \frac{\#\text{ correct predictions}}{\#\text{ test instances}}$$

$$\text{error} = 1 - \text{accuracy} = \frac{\#\text{ incorrect predictions}}{\#\text{ test instances}}$$

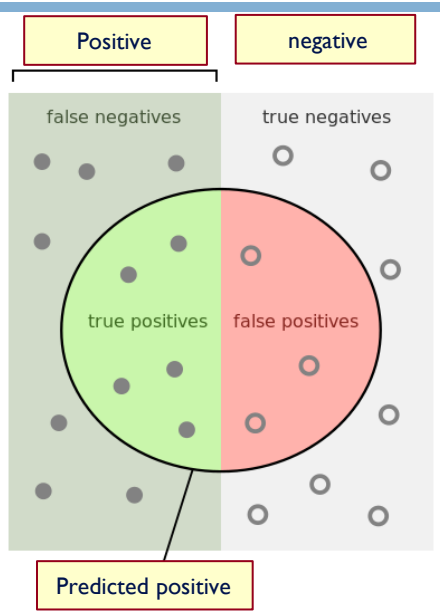
- Any possible problems with it?

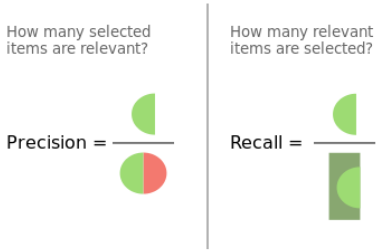
Alternative Metrics

- If the Binary classification problem is biased
  - In many problems most examples are negative
- Or, in multiclass classification
  - The distribution over labels is often non-uniform
- Simple accuracy is not a useful metric.
  - Often we resort to task specific metrics
- However one important example that is being used often involves Recall and Precision

- **Recall:** $\frac{\text{\# (positive identified = true positives)}}{\text{\# (all positive)}}$

- **Precision:** $\frac{\text{\# (positive identified = true positives)}}{\text{\# (predicted positive)}}$

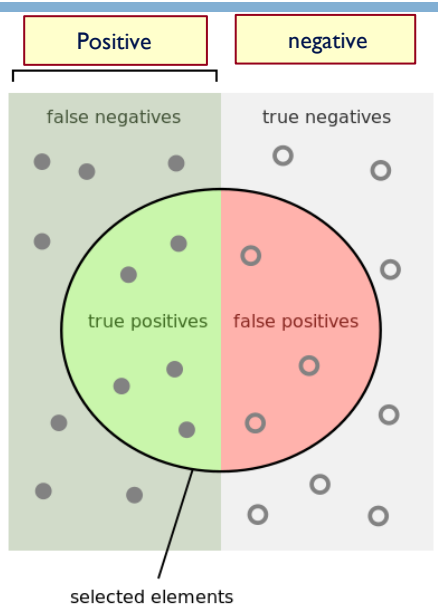


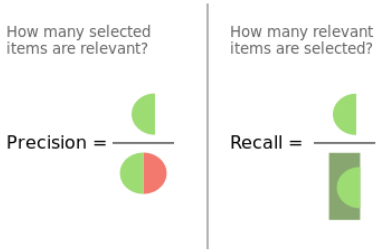


Example

- 100 examples, 5% are positive.
- Just say NO: your accuracy is 95%
  - Recall = precision = 0
- Predict 4+, 96-; 2 of the +s are indeed positive
  - Recall:2/5; Precision: 2/4

- **Recall:** $\frac{\#\text{ (positive identified = true positives)}}{\#\text{ (all positive)}}$
- **Precision:** $\frac{\#\text{ (positive identified = true positives)}}{\#\text{ (predicted positive)}}$

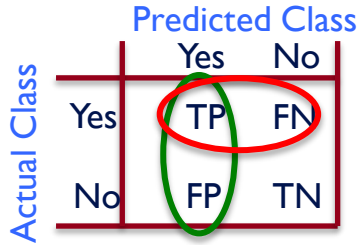




Confusion Matrix

The notion of a confusion matrix can be usefully extended to the multiclass case (i,j) cell indicate how many of the i-labeled examples were predicted to be j

- Given a dataset of P positive instances and N negative instances:



$$\text{accuracy} = \frac{TP + TN}{P + N}$$

- Imagine using classifier to identify positive cases (i.e., for information retrieval)

$$\text{precision} = \frac{TP}{TP + FP}$$

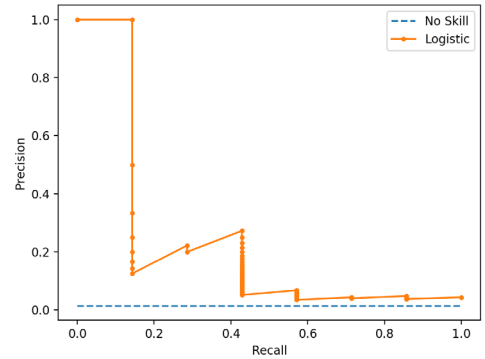
Probability that a randomly selected positive prediction is indeed positive

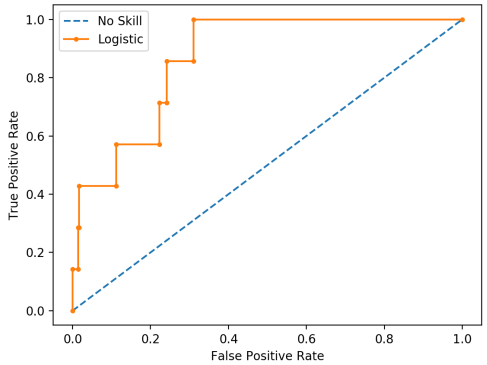
$$\text{recall} = \frac{TP}{TP + FN}$$

Probability that a randomly selected positive is identified

Relevant Metrics

- Recall-Precision Curve: Plot of Recall (x) vs Precision (y).
- ROC Curve: Plot of False Positive Rate (x) vs. True Positive Rate (y).
- It makes sense to consider Recall and Precision together or combine them into a single metric.
- AUC – Area Under the Curve





Relevant Metrics

- F-Measure:
  - A measure that combines precision and recall is the harmonic mean of precision and recall.
  - F1 is the most commonly used metric.

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{\beta^2 \cdot precision + recall}$$

Metrics

Methodologies

Statistical Significance

Penn Engineering

Comparing Classifiers

Say we have two classifiers, *C1* and *C2*, and want to choose the best one to use for future predictions

Can we use training accuracy to choose between them?

- No!

- What about accuracy on test data?
- Yes, but…
  - We basically want to look at more than a single number; gather some statistical evidence.

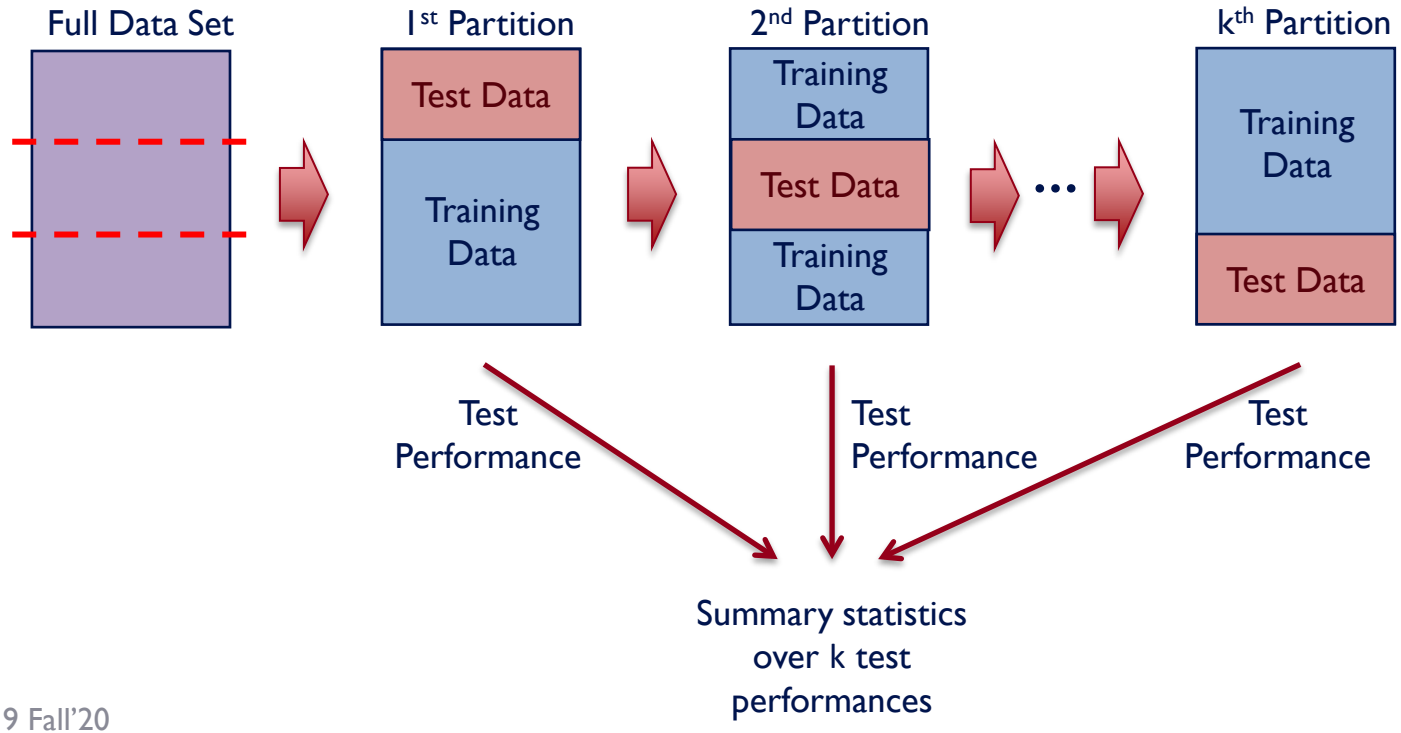
k-fold cross validation

- Instead of a single test-training split:

- Split data into k equal-sized parts

- Train and test k different classifiers
- Report average accuracy and standard deviation of the accuracy

Example 3-Fold CV



More on Cross-Validation

- Cross-validation generates an approximate estimate of how well the classifier will do on "unseen" data
  - As $k \rightarrow n$, the model becomes more accurate (more training data)
    ...but, CV becomes more computationally expensive.
  - Choosing $k < n$ is a compromise. $k = 5$ is often used.
  - $k = n$ is called "leave-one-out";

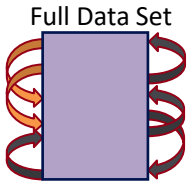
- Averaging over different partitions is more robust than just a single train/validate partition of the data

- It is an even better idea to do CV repeatedly!

Multiple Trials of k-Fold CV

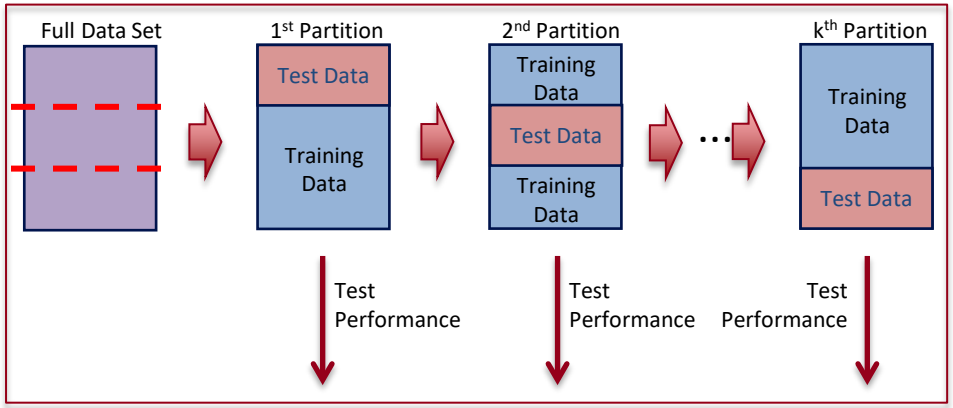
1.) Loop for t trials:

a.) Randomize Data Set



Shuffle

b.) Perform k-fold CV

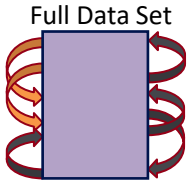


2.) Compute statistics over $t \times k$ test performances

Comparing Multiple Classifiers

1.) Loop for t trials:

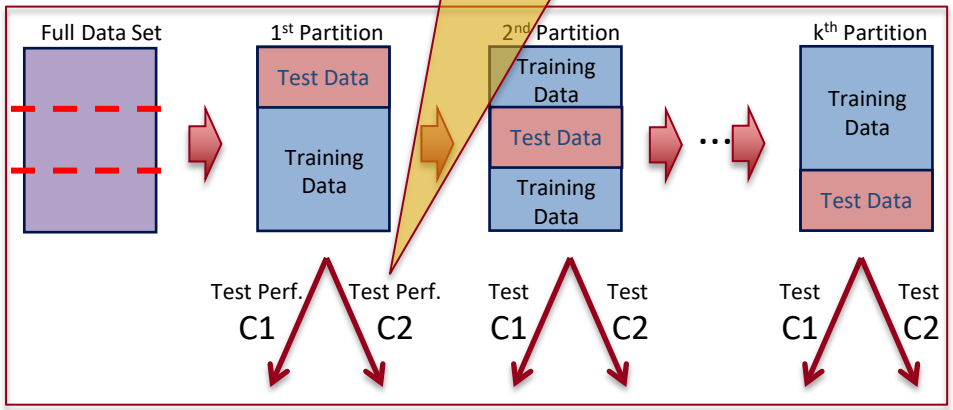
a.) Randomize Data Set



Shuffle

Test each candidate learner on same training/testing splits

b.) Perform k-fold CV

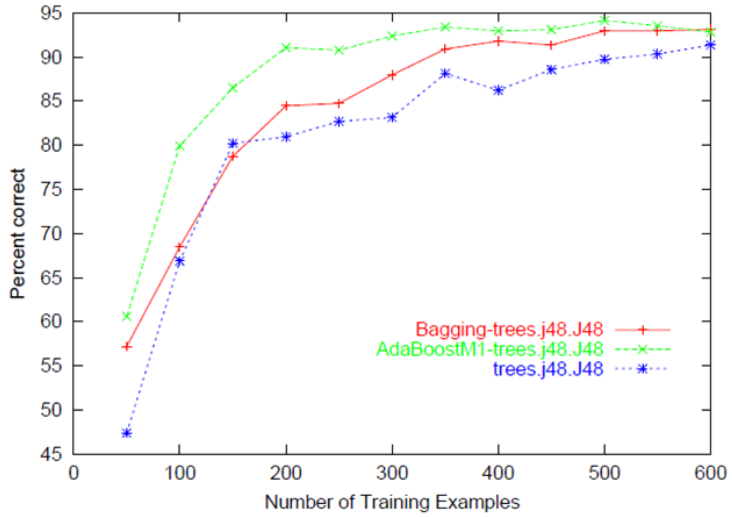


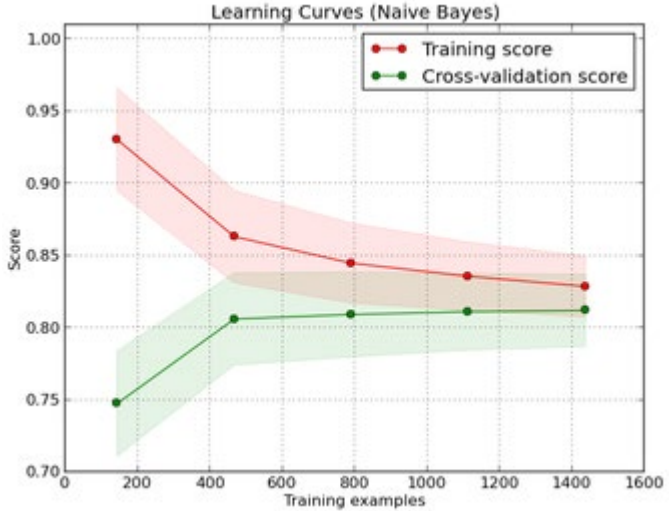
2.) Compute statistics over $t \times k$ test performances

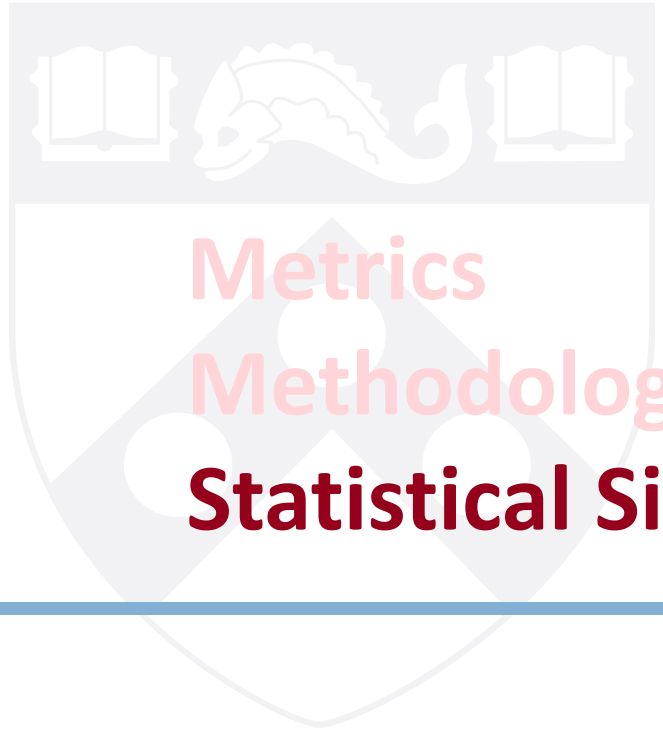
Allows us to do paired summary statistics (e.g., paired t-test)

Learning Curve

- Shows performance versus the # training examples
  - Compute over a single training/testing split
  - Then, average across multiple trials of CV







Metrics

Methodologies

Statistical Significance

Penn Engineering

Evaluation: Significance Tests

- You have two different classifiers, A and B
- You train and test them on the same data set using N-fold cross-validation
- For the n-th fold:

  accuracy(A, n), accuracy(B, n)

  p_n = accuracy(A, n) - accuracy(B, n)
- Is the difference between A and B's accuracies significant?

Hypothesis testing

- [Next we are introducing a methodology for answering the question: can we distinguish two models? Which one is better?]
- You want to show that hypothesis H is true, based on your data
  - (e.g. H = "classifier A and B are different" or "classifier A is better than B")
- Define a null hypothesis H_0 and an alternative hypothesis H_1
  - (H_0 is the contrary of what you want to show, it is the default position you want to prove wrong)
- Decide on some statistic M
  - e.g. the difference in accuracy between A and B given that H_0 is true
- Describe the hypotheses in terms of the test statistic
  - e.g. H_0: $M = 0, H_1: M \neq 0$
- Can you refute (reject) H_0?

Rejecting H_0

- H_0 defines a distribution $P(M \mid H_0)$ over the statistic M
- Select a significance value α – the probability of rejecting the null hypothesis when it is actually true
  - (e.g. 0.05, 0.01, etc.)
- Compute the observed value of the test statistic M_{obs} from your data
  - e.g. the average difference in accuracy over your N folds
- The distribution of the test statistic under the null hypothesis partitions the possible values of M into those for which the null hypothesis is rejected.
- Refute H_0 if M_{obs} is inside the "critical zone".

Rejecting H_0 – Another Alternative

- H_0 defines a distribution $P(M \mid H_0)$ over the statistic M
- Select a significance value α
  - (e.g. 0.05, 0.01, etc.)
- Compute the observed value of the test statistic M_{obs} from your data
  - e.g. the average difference in accuracy over your N folds
- Compute $P(M \geq M_{obs} | H_0)$ – called the p-value
- Refute H_0 with a significance level α if $P(M \geq M_{obs} | H_0) \leq \alpha$

Statistical Concepts - Example

- Say we wish to know if getting up from the chair and exercise after today's class will help us to lose weight.
- The null hypothesis: exercise does not effect weight loss
- The alternative hypothesis: exercise does have an effect!
- We take a poll and ask people of they have lost weight after exercise
- The test statistic could be – the number of people who did lose weight, or the amount of weight that they have lost.

Statistical Concepts - Example

- The test statistic depends on how exactly we formulate the hypotheses (w.r.t. the total weight of all people or something else).
- The p-value could be something like: how probable is it to see a wight loss of 10 pounds or more when remaining seated on the chair?
- If this probability is very low, then we will reject the null hypothesis and conclude that **after class** we should exercise.

Metrics

Methodologies

Statistical Significance

Paired t-test

McNemar

Bootstrap

Penn Engineering

Metrics

Methodologies

Statistical Significance

Paired t-test

McNemar

Bootstrap

Penn Engineering

Paired t-test

- A paired t-test is used to compare two **population means** where you have two samples in which observations in one sample can be paired with observations in the other sample.

Paired t-test

- Paired tests produce tighter bounds since any difference is due to difference in the hypotheses rather than differences in the test set.
- Significance Testing of the Paired Tests:
- Compute the statistics:

$$t = \frac{\delta\sqrt{k}}{\sqrt{\frac{1}{k-1}\sum_{i=1}^{k}(\delta_i - \delta)^2}}$$

- where δ_i is the measured difference between A and B on the $i-th$ data set and $\delta = \frac{1}{k}\sum_i \delta_i$ is their average.
- The statistics is distributed according to a t-distribution(k)

Paired t-test

- Null hypothesis (H_0; to be refuted):
  - There is no difference between A and B, i.e. the expected accuracies of A and B are the same
- That is, the expected difference (over all possible data sets) between their accuracies is 0:

  H_0: $E[p_D] = 0$
- We don't know the true $E[p_D]$
- *N*-fold cross-validation gives us *N* samples of p_D

Paired t-test

- Null hypothesis H_0: $E[diff_D] = \mu = 0$
- *m:* our estimate of μ based on *N* samples of $diff_D$

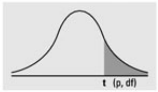
  $m = 1/N \sum_n diff_n$
- The estimated variance S^2:

  $S^2 = 1/(N-1) \sum_{1,N} (diff_n - m)^2$
- Accept Null hypothesis at significance level α if the following statistic lies in $(-t_{\alpha/2, N-1}, +t_{\alpha/2, N-1})$

$$\frac{\sqrt{N}m}{S} \sim t_{N-1}$$

T – Distribution Table

Numbers in each row of the table are values on a *t*-distribution with (*df*) degrees of freedom for selected right-tail (greater-than) probabilities (*p*).



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905
CI	———	———	80%	90%	95%	98%	99%	99.9%

Paired t-test example

- Question: The downtimes (measured in hours) for computer systems in six branches of a major bank were recorded for year 1 and year 2. Compute the test statistics for the paired t-test.
- Solution:

Branch	Year 1	Year 2	Difference (Year 1 – Year 2)	Square of Difference
A	40	30	10	100
B	54	41	13	169
C	32	24	8	64
D	36	38	-2	4
E	55	56	-1	1
F	46	37	9	81
			Sum = 37	**Sum = 419**

Paired t-test

- Sample size: $n = 6$
- Sum of differences $\sum d_i = 37$
- Sum of squared differences $\sum d_i{}^2 = 419$
- Mean of case-wise differences: $\bar{d} = \frac{\sum d_i}{n} = \frac{37}{6} = 6.166$
- Standard Deviation : $s_d = \frac{\sum d_i - n\bar{d}^2}{n-1} = \sqrt{\frac{419 - 6 \times (6.166)^2}{5}} = 6.177$
- Test statistics for the paired t-test: $t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} = 2.445$

Metrics

Methodologies

Statistical Significance

Paired t-test

McNemar

Bootstrap

Penn Engineering

McNemar's Test

The test is often used for the situation where one tests for the presence (1) or absence (0) of something and variable A is the state at the first observation (i.e., pretest) and variable B is the state at the second observation (i.e., post-test).

- An alternative to Cross Validation, when the test can be run only once, but you have **access to predictions on individual examples**.

McNemar's Test

- Divide the sample S into a training set R and a test set T.
- Train algorithms A and B on R, yielding classifiers A, B
- Record how each example in T is classified and fill the following table:

# of examples misclassified by both A and B N_{00}	# of examples misclassified by A but not B N_{01}
# of examples misclassified by B but not A N_{10}	# of examples misclassified by neither A nor B N_{11}

where N is the total number of examples in the test set T

$$N_{00} + N_{10} + N_{01} + N_{11} = N$$

McNemar's Test

- The null hypothesis: the two learning algorithms have the same error rate on a randomly drawn sample. That is, we expect that

$$N_{10} = N_{01}$$

- The statistics we use to measure deviation from the expected counts:

$$\frac{(|N_{01} - N_{10}| - 1)^2}{N_{01} + N_{10}}$$

- This statistics is distributed (approximately) as t-distribution with 1 degree of freedom

Which model has higher accuracy?

Scenario A		Scenario B	
$N_{00} = 29$	$N_{01} = 1$	$N_{00} = 15$	$N_{01} = 15$
$N_{10} = 11$	$N_{11} = 9959$	$N_{10} = 25$	$N_{11} = 9945$

Model 1 in Scenario A and in Scenario B

Model 2 in Scenario A and in Scenario B

Model 1 in Scenario A and model 2 in Scenario B

Model 2 in Scenario A and model 1 in Scenario B

None of the above

What are the hypotheses?

H0: $N\_10 = N\_01$ H1: N_10 != N_01

H0: N_10 != N_01
H1: N_10 != N_01

H0: N_10 > N_01
H1: N_10 <= N_01

H0: N_10 < N_01
H1: N_10 >= N_01

None of the above

McNemar's Test - Example

Scenario A		Scenario B	
$N_{00} = 29$	$N_{01} = 1$	$N_{00} = 15$	$N_{01} = 15$
$N_{10} = 11$	$N_{11} = 9959$	$N_{10} = 25$	$N_{11} = 9945$

- Now let's calculate the test statistic for both scenarios:

$$\frac{(|N_{01} - N_{10}| - 1)^2}{N_{01} + N_{10}}$$

$$\chi_A^2 = 6.75 \qquad \chi_B^2 = 2.025$$

- We get $p - val_A = 0.0093$ and $p - val_B = 0.15$

Metrics

Methodologies

Statistical Significance

Paired t-test

McNemar

Bootstrap

Penn Engineering

Bootstrap Hypothesis Testing

Sometimes, we are not interested in comparing the **mean** performance of the two classifiers.

When comparing different statistics, we cannot use the t-test because we do not know the distribution of the test statistic under the null hypothesis.

In this case we can use statistical tests that are called non-parametric tests. One of them is Bootstrap.

Bootstrap Hypothesis Testing

- Another alternative to Cross Validation, when the test can be run only once, but you have **access to predictions on individual examples**.
- Divide the sample S into a training set R and a test set T.
- Train algorithms A and B on R, yielding classifiers A, B.
- Record how each example in T is classified.
- Draw multiple samples from our original sample **with replacements**.

Bootstrap Hypothesis Testing

- Take M samples $X_1^*, \ldots, X_N^*$ from the original test set performance values (e.g., the differences in performance between the classifiers on every example in the test set) $X_1, \ldots, X_N$ **with replacements** and calculate the test statistic from each sample:

$$X_{(1)}^{*(1)}, \ldots, X_{(N)}^{*(1)} \rightarrow t_{(1)}^*$$
$$\vdots$$
$$X_{(1)}^{*(M)}, \ldots, X_{(N)}^{*(M)} \rightarrow t_{(M)}^*$$

- Calculate a test statistic for each sample $t_{(1)}^*, \ldots, t_{(M)}^*$
- Use the following formula to calculate the p-value:

$$p = \frac{\#\{t_{(i)}^* \geq t\}}{M}$$