Model Evaluation

AW

Lecture Overview

1. **Missing Attributes Handling**
2. **Model Evaluation**
3. **Performance Comparison**

Handling Missing Attribute Values

- Missing values affect decision tree construction in three different ways:
  - Affects how impurity measures are computed
  - Needs method of how to distribute instance with missing value to child nodes
  - Needs method of how a test instance with missing value is classified

Computing Impurity Measure

Tid	Refund	Marital Status	Taxable Income	Class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	?	Married	90K	Yes

Missing value

Before Splitting:

$$Entropy(Parent)$$
$$= -0.3\log(0.3) - (0.7)\log(0.7)$$
$$= 0.8813$$

	Class = Yes	Class = No
Refund=Yes	0	3
Refund=No	2	4
Refund=?	1	0

Split on Refund:

$$Entropy(Refund = Yes) = 0$$

$$Entropy(Refund = No)$$
$$= -(2/6)\log(2/6) - (4/6)\log(4/6)$$
$$= 0.9183$$

$$Entropy(Children)$$
$$= 0.3\,(0) + 0.6\,(0.9183) = 0.551$$

$$Gain = 0.9 \quad 0.8813 - 0.551 = 0.2417$$

Distribute Instances-J48 solutions

Tid	Refund	Marital Status	Taxable Income	Class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No

Refund

Yes

Class=Yes	**0**
Class=No	**3**

No

Cheat=Yes	**2**
Cheat=No	**4**

Tid	Refund	Marital Status	Taxable Income	Class
10	?	Single	90K	Yes

Refund

Yes

Class=Yes	**0 + 3/9**
Class=No	**3**

No

Class=Yes	**2 + 6/9**
Class=No	**4**

Probability that Refund=Yes is 3/9

Probability that Refund=No is 6/9

Assign record to the left child with weight = 3/9 and to the right child with weight = 6/9

Build DT With Distributed Instances

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Class**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	**?**	Married	90K	**Yes**

Refund

Yes

Class=Yes	**3/9**
Class=No	**3**

No

Marrital Status

Single, Divorced

Class=Yes	**2**
Class=No	**1**

Married

Class=Yes	**6/9**
Class=No	**3**

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Class**
3	No	Single	70K	**No**
5	No	Divorced	95K	**Yes**
8	No	Single	85K	**Yes**

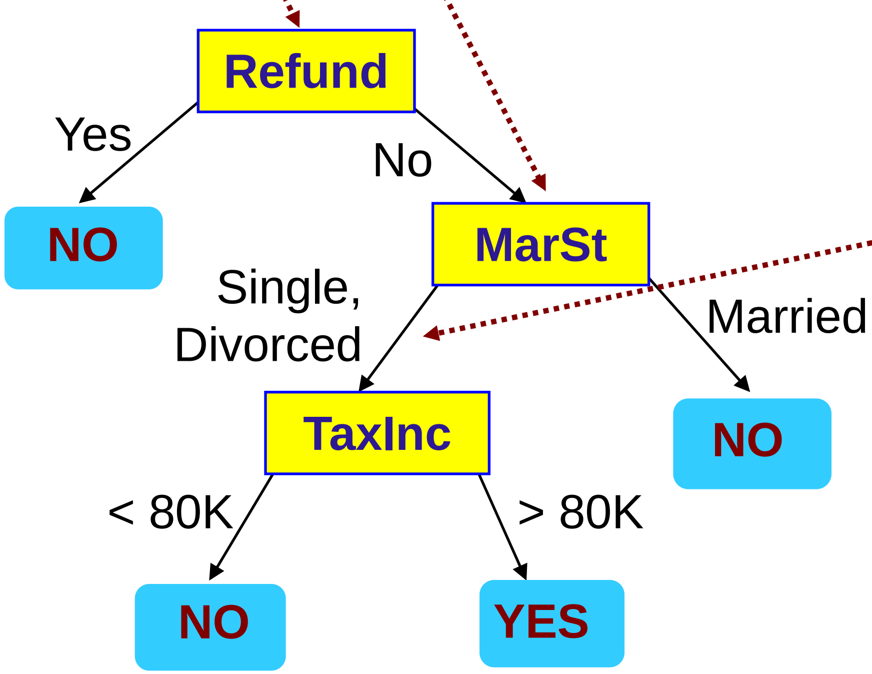
Tid	**Refund**	**Marital Status**	**Taxable Income**	**Class**
2	No	Married	100K	**No**
6	No	Married	60K	**No**
9	No	Married	75K	**No**
10	**?**	Married	90K	**Yes**

Classify New Instances w/Missing Data

New record:

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Class**
11	No	?	85K	?

	Married	Single	Div.	Total
Class=No	3	1	0	4
Class=Yes	6/9	1	1	2.67
Total	3.67	2	1	6.67



$$\Pr(Marital\ Status = Married) = \frac{3.67}{6.67}$$

$$\Pr(Marital\ Status \in \{Single, Divorced\}) = \frac{3}{6.67}$$

Lecture Overview

1. Missing Attributes Handling
2. Model Evaluation
3. Performance Comparison

Data Fragmentation

- Number of instances gets smaller as you traverse down the tree
- Number of instances at the leaf nodes could be too small to make any statistically significant decision

Model Evaluation

- Metrics for Model Evaluation
  - How to evaluate the performance of a model?
- Methods for Model Evaluation
  - How to obtain reliable estimates?
- Methods for Model Comparison
  - How to compare the relative performance among competing models?

Metrics for Performance Evaluation

- Focus on the predictive capability of a model
  - Rather than how fast it takes to classify or build models, scalability, etc.
- Confusion Matrix:

	PREDICTED CLASS		
		Class = Yes	Class = No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

Metrics for Performance Evaluation...

-

		PREDICTED CLASS	
	sed metric:	Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

Limitation of Accuracy

- Consider a 2-class problem
  - Number of Class 0 examples = 9990
  - Number of Class 1 examples = 10

- If model predicts everything to be class 0, accuracy is 9990/10000 = 99.9 %
  - Accuracy is misleading because model does not detect any class 1 example

Cost Matrix

	PREDICTED CLASS		
ACTUAL CLASS	C(i|j)	**Class=Yes**	**Class=No**
	Class=Yes	C(Yes|Yes)	C(No|Yes)
	Class=No	C(Yes|No)	C(No|No)

$C(i|j)$: Cost of misclassifying class j example as class i

Computing Cost of Classification

Cost Matrix	PREDICTED CLASS		
	C(i|j)	+	-
ACTUAL CLASS	+	-1	100
	-	1	0

Model M_1	PREDICTED CLASS		
		+	-
ACTUAL CLASS	+	150	40
	-	60	250

Accuracy = 80%

Cost = 3910

Model M_2	PREDICTED CLASS		
		+	-
ACTUAL CLASS	+	250	45
	-	5	200

Accuracy = 90%

Cost = 4255

Cost vs. Accuracy

Count	PREDICTED CLASS	
ACTUAL CLASS	Class=Yes	Class=No
Class=Yes	a	b
Class=No	c	d

Accuracy is proportional to cost if we have

1. C(Yes|No)=C(No|Yes) = qC
2. C(Yes|Yes)=C(No|No) = pC

and $p = 1 - q$

Let $N = a + b + c + d$

then $Accuracy = (a + d)/N$

Cost	PREDICTED CLASS	
ACTUAL CLASS	Class=Yes	Class=No
Class=Yes	pC	qC
Class=No	qC	pC

$$
\begin{aligned}
Cost &= pC\,(a + d) + qC\,(b + c) \\
&= pC(a + d) + qC(N - a - d) \\
&= qCN - (q - p)C(a + d) \\
&= NC[q - (q - p) \times Accuracy]
\end{aligned}
$$

Cost-Sensitive Measures

$$Precision(p) = \frac{a}{a + c}$$

$$Recall\,(r) = \frac{a}{a + b}$$

$$F\text{ - }measure(F) = \frac{2rp}{r + p} = \frac{2a}{2a + b + c}$$

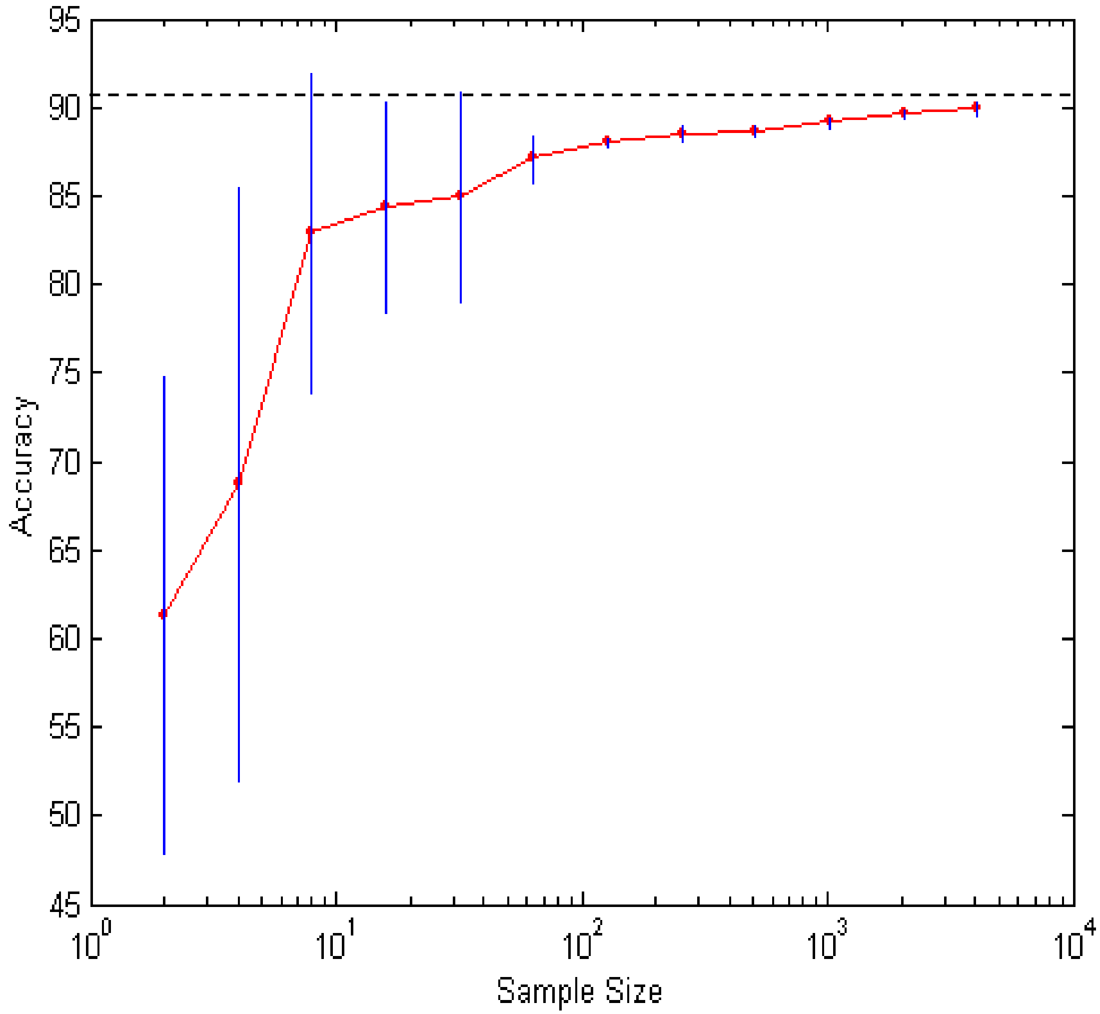
- Precision is biased towards C(Yes|Yes) & C(Yes|No)
- Recall is biased towards C(Yes|Yes) & C(No|Yes)
- F-measure is biased towards all except C(No|No)

$$Weighted\ Accuracy = \frac{w_1 a + w_4 d}{w_1 a + w_2 b + w_3 c + w_4 d}$$

Methods for Performance Evaluation

- How to obtain a reliable estimate of performance?
- Performance of a model may depend on other factors besides the learning algorithm:
  - Class distribution
  - Cost of misclassification
  - Size of training and test sets

Learning Curve



Learning curve shows how accuracy changes with varying sample size

Requires a sampling schedule for creating learning curve:

- Arithmetic sampling (Langley, et al)
- Geometric sampling (Provost et al)

Effect of small sample size:

- Bias in the estimate
- Variance of estimate

Arithmetic and Geometric Sampling

- Both methods are a progressive sampling that increase sample size until the objective is satisfied
- Objective = stopping condition for both methods is when the DT measure of accuracy stop changing (measure-wise $m(i) = m(i+1)$).
- Arithmetic sampling : sizes are increased so that sample sizes form an arithmetical progression, i.e. $S_i = S_0 + i * C$ where C is a constant and S_0 initial sample size
- Geometric sampling: sizes are increased so that sample sizes form a geometric progression, i.e. $S_i = S_0 * C \cdot i$, where S_0 and C are as before

Methods of Estimation

- Holdout
  - Reserve 2/3 for training and 1/3 for testing
- Random subsampling
  - Repeated holdout
- Cross validation
  - Partition data into k disjoint subsets
  - k-fold: train on $k-1$ partitions, test on the remaining one
  - Leave-one-out: $k = n$
- Stratified sampling
  - oversampling vs undersampling
- Bootstrap
  - Sampling with replacement
  - Records that are not chosen become test set.
  - Since probability of being chosen uniformly at random is
  - $1-\left(1-\frac{1}{n}\right)^n \sim 1-e^{-1}$ on the average ~0.633 of the records are chosen

Lecture Overview

1. Missing Attributes Handling
2. Model Evaluation
3. Performance Comparison

Comapring 2 Models: Test of Significance

- Given two models:
  - Model M1: accuracy = 85%, tested on 30 instances
  - Model M2: accuracy = 75%, tested on 5000 instances
- Can we say M1 is better than M2?
  - How much confidence can we place on accuracy of M1 and M2?
  - Can the difference in performance measure be explained as a result of random fluctuations in the test set?

Confidence Interval for Accuracy

- Recall: prediction can be regarded as a Bernoulli trial
  - A Bernoulli trial has 2 possible outcomes
  - Possible outcomes for prediction: correct or wrong
  - Collection of Bernoulli trials has a Binomial distribution:
    - $x \sim Bin(N, p)$ where x: number of correct predictions
    - e.g: Toss a fair coin 50 times, how many heads would turn up?

      Expected number of heads $= N \times p = 50 \times 0.5 = 25$

      Variance of the number of heads$= N \times p \times (1-p)$

      $=25 \times 0.5 = 12.5$
- Given x (# of correct predictions) or equivalently, $acc = x/N$, and N (# of test instances). Can we predict p (true accuracy of model)?

Comparing Performance of 2 Models

- Given two models, say M_1 and M_2, which is better?
  - M_1 is tested on D_1 (size=n_1), found x_1 incorrect predictions, experimental error rate = e_1
  - M_2 is tested on D_2 (size=n_2), found x_2 incorrect predictions, experimental error rate = e_2
  - Bernoulli trials – x_i - is random Binomially distributed number of errors in model $M_i \Rightarrow x_i/n_i$ is random binomially distributed error rate in model M_i after n_i trials. Theoretically

$$\mu_i = E(x_i)/n_i = n_i \times e_i/n_i = e_i$$

and

$$\sigma_i^2 = E((x_i/n_i - E(x_i/n_i))^2) = \mu_i \times (1-\mu_i)/n_i$$

  - If n_i is big enough approximate binomial distribution with normal

$$e_1 \sim N(\mu_1, \sigma_1) \qquad e_2 \sim N(\mu_2, \sigma_2)$$

  - Estimate

$$\hat{\mu}_i = e_i \text{ and } \hat{\sigma}_i = \frac{e_i(1-e_i)}{n_i}$$

What are We Testing?

- We assume that we know theoretical distribution of error rate (binomial modeled by normal), so we have accurate estimates of its variance, without using statistic (i.e. not sample variance)!
- Difference in accuracy is a random variable. Since $D1$ and $D2$ are independent, difference in accuracy is normally distributed variable with mean $\mu_1 - \mu_2$ and $\sigma_1{}^2 - \sigma_2{}^2$
- We test whether the sample mean difference in accuracy that we know ($d = e_1 - e_2$) is different from the unknown theoretical difference in accuracy.
- Hypothesis that we are testing is $H0\colon \mu_1 - \mu_2 = 0$. So we pick the level of confidence at which we expect $H0$ to hold.
- Then we find the interval in which the theoretical mean difference lies around d at a given confidence level
- If it includes 0 then $H0$ is indeed valid.

Comparing Performance of 2 Models

- To test if performance difference is statistically significant:
  $d = e_1 - e_2$
  - $d \sim N(d_t, \sigma_t)$ where d_t is the true difference
  - Since $D1$ and $D2$ are independent, their variance adds up:

$$\boldsymbol{\sigma}_t^2 = \boldsymbol{\sigma}_1^2 + \boldsymbol{\sigma}_2^2 \cong \hat{\boldsymbol{\sigma}}^2{}_t = \hat{\boldsymbol{\sigma}}_1^2 + \hat{\boldsymbol{\sigma}}_2^2$$
$$= \frac{e_1(1-e_1)}{n_1} + \frac{e_2(1-e_2)}{n_2}$$

  - When $n_1 = n_2 = N$ we have $\sigma_t^2 = \frac{e_1(1-e_1)+e_2(1-e_2)}{N}$
  - At $(1-\alpha)$ confidence level,

$$d_t = d \pm Z_{\alpha/2}\hat{\boldsymbol{\sigma}}_t$$

An Illustrative Example

- Given: M_1: $n_1 = 30, e_1 = 0.15$
  M_2: $n_2 = 5000, e_2 = 0.25$
- $d = |e_2 - e_1| = 0.1$ (2-sided test)
- At 95% confidence level, $Z_{\alpha/2}=1.96$

$$\hat{\sigma}_d^{\,2} = \frac{0.15(1-0.15)}{30} + \frac{0.25(1-0.25)}{5000} = 0.0043$$

$$d_t = 0.100 \pm 1.96 \times \sqrt{0.0043} = 0.100 \pm 0.128$$

=> Interval contains 0 => difference may not be statistically significant

An Illustrative Example (cont.)

- At what confidence the difference is statistically significant?

$$\hat{\sigma}^2{}_d = 0.0043$$

$$d_t = 0.100 \pm Z_{\alpha/2} \times \sqrt{0.0043} > 0\,?$$

$$Z_{\alpha/2} < \frac{0.100}{\sqrt{0.0043}} = 1.527 \Rightarrow (1-\alpha) = .937$$

=> Interval does not contains 0 => difference may is statistically significant

Comparing Performance of 2 Algorithms

- Each learning algorithm may produce k models:
  - L_1 may produce $M_{11}, M_{12}, \ldots, M_{1k}$
  - L_2 may produce $M_{21}, M_{22}, \ldots, M_{2k}$
- If models are generated on the same test sets $D_1, D_2, \ldots, D_k$ (e.g., via k-cross-validation)
  - For each test set: compute $d_j = e_{1j} - e_{2j}$
  - d_j are i.i.d. variables with mean d_t and variance σ_t
  - Estimate d_t with $\bar{d} = \frac{1}{k}\sum_1^k d_i$
  - So what is the variance then? we do not know it, so we need to estimate it statistically

Confidence Intervals - Student Distribution

So what are we doing with these confidence interval calculations?

- We are trying to determine whether the true *unknown* mean, d_t is based on a sample mean $\bar{d}$. It's a range of plausible values for d_t.
- But the calculation of confidence intervals (so far) assumes we know true standard deviation σ *(or at least have a good estimate of it without using statistic)*
- This doesn't often happen in real life. If we are trying to estimate μ, we will also probably have to estimate σ *for unknown distribution from the same statistic.*
- What's our sample-based estimate of the standard deviation? s
- This throws off everything. The calculation is no longer based on a normal distribution, but a t-distribution

Confidence Intervals - Student (cont.)

- When the true standard deviation σ is not known we need to use s instead. The usual formula:

$$\left(\bar{x} - z^* \frac{\sigma}{\sqrt{n}}, \bar{x} + z^* \frac{\sigma}{\sqrt{n}}\right)$$

is then replaced by:

$$\left(\bar{x} - t^*_{df} \frac{s}{\sqrt{n}}, \bar{x} + t^*_{df} \frac{s}{\sqrt{n}}\right)$$

where the multiplier z^* is replaced by a value from the t-distribution with $df = (n - 1)$ `degrees of freedom'.

- the t-distribution is really a family of distributions that look like the normal distribution, but is spread out further (fatter tails).

Confidence Intervals - Student (cont.)

TABLE D *t* distribution critical values

	Upper tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

Significance Test for Unknown Variance

- To test the hypothesis H0: $\mu = \mu_0$ against an alternative hypothesis, compute the one-sample t-statistic

$$t_{df} = \frac{\bar{x} - \mu_0}{s / \sqrt{N}}$$

computed using sample mean and variance

- Confidence p-values are computed by comparing the statistic with a t-distribution with $df = N - 1$.

2 Algorithms Peformance Comparison (cont.)

- Estimate:

$$s_t^2 = \frac{\sum_{j=1}^{k}(d_j - \bar{d})^2}{(k-1)};$$

- Hypothesis $H0$ that we are testing is that $d_t = 0$ so we find confidence interval

$$d_t = d \pm t_{1-\alpha,k-1} \frac{s_t}{\sqrt{k}}$$

at appropriate confidence level and if contains 0 then $H0$ is valid

Instructive Example

In 7-cross validation experiments on two DT trees out of 2000 test records accurately were evaluated

DT_1: 1200 1219 1103 1213 1258 1325 1295

DT_2: 1247 1098 1185 1087 1377 1363 1121

- So $d_1, \dots d_7$ values are:

  -0.024 0.061 -0.041 0.063 -0.060 -0.019 0.087

- $\bar{d}$=0.0095, s=0.0589; at 95% level of confidence

  $0.0095-2.365\times 0.0589/\sqrt{7} \le d_t \le 0.0095+2.365\times 0.0589/\sqrt{7}$ or $-0.042 \le d_t \le 0.061$ so the difference between models is not statistically significant

Reading

- 2.2, 3.6-3.8