

Decision Trees: Evaluation

Mauro Sozio*

*slides adapted from the course: Introduction to data mining, Steinbach, Kumar

Model Evaluation

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

Model Evaluation

- | Metrics for Performance Evaluation
 - How to evaluate the performance of a model?
- | Methods for Performance 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 running time, scalability, etc.
- | Confusion Matrix:

ACTUAL Class	PREDICTED Class		
		Class=Yes	Class=No
	Class=Yes	TP	FN
	Class=No	FP	TN

TP (true positive)
FN (false negative)
FP (false positive)
TN (true negative)

Metrics for Performance Evaluation...

	PREDICTED Class		
		Class=Yes	Class=No
	Class=Yes	TP	FN
	Class=No	FP	TN

Text

- | Most widely-used metric:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

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 $9990/10000 = 99.9\%$
 - Misleading: model does not detect any Class 1 example

Cost Matrix and Weighted Accuracy

$C(i|j)$: Cost of misclassifying Class j record as Class i

	PREDICTED Class		
	$C(i j)$	Class=Yes	Class=No
ACTUAL Class	Class=Yes	$C(\text{Yes} \text{Yes})$	$C(\text{No} \text{Yes})$
	Class=No	$C(\text{Yes} \text{No})$	$C(\text{No} \text{No})$

$$\text{Weighted Acc.} = \frac{\text{TP} \times C(\text{Yes}|\text{YES}) + \text{TN} \times C(\text{No}|\text{No})}{\text{TP} \times C(\text{Yes}|\text{YES}) + \text{FN} \times C(\text{No}|\text{yes}) + \text{FP} \times C(\text{Yes}|\text{No}) + \text{TN} \times C(\text{No}|\text{No})}$$

Cost of Classification

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

Cost Matrix Penalizing FN

$$Cost = TP \times c(TP) + TN \times c(TN) + FP \times c(FP) + FN \times c(FN)$$

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

Accuracy = $400 / 500 = 80\%$
Cost = 3910

Model M_2	PREDICTED Class		
		+	-
	ACTUAL Class	+	-
		205	65
		25	200

Accuracy = $405 / 500 = 81\%$
Cost = 6320

Precision, Recall, F-measure

$$\text{Precision (p)} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall (r)} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Easy to get high precision: ‘Yes’ only for instances we are “sure” about (e.g. easy to classify “robot” if 1000 pages/sec. retrieved) => low recall

Trivial to get high recall: classify all instances as ‘Yes’
=> low precision

Precision, Recall, F-measure

F-measure is the harmonic mean of p and r, which penalizes very small recall or very small precision

$$\text{F-measure (F)} = \frac{2rp}{r+p} \quad (\text{Harmonic Mean of p and r})$$

E.g. : Harmonic Mean (1,2,4) = $1 / (1/3 (1+1/2+1/4)) = 12/7$
(reciprocal of the arithmetic mean of the reciprocals)

It can be shown $\min (x_1, \dots, x_n) \leq H(x_1, \dots, x_n) \leq n * \min (x_1, \dots, x_n)$
 \Rightarrow min is very small then harmonic mean is very small.

Model Evaluation

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

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

Methods of Estimation

- | Holdout
 - Reserve 2/3 for training and 1/3 for testing
- | Random subsampling
 - Repeated holdout (compute the average)
- | Cross validation
 - Partition data into k disjoint subsets of the same size
 - k -fold: train on $k-1$ sets, test on the remaining one. Repeat k times, each subset being used exactly once for testing. Compute the average of the results.
 - Leave-one-out: $k=n$, (1 record training test, rest test set)
- | Bootstrap
 - Sampling with replacement

Model Evaluation

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

Test of Significance

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

- | 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, M2?

Confidence Interval for Accuracy

- | Estimate accuracy by running classifier on all possible instances, **unfeasible!**
- | If we have a sample of all records (uniform and independent), then, determining CI for accuracy is related to the problem:
 - *we are given a coin with p being the probability of getting a head. We flip it N times, getting x heads.*
 - ◆ is $p' = x/N$ close to p ?
 - ◆ how much confidence on the accuracy of p' ?

Bernoulli Experiment

- | Given a coin with probability p of getting a 'head', flipped N times. Let X be the random variable denoting the number of heads.
- | $\text{Prob}(X = x) = \binom{N}{p} p^x (1-p)^{N-x}$
- | Example: coin is fair ($p=0.5$), flipped 50 times, then the prob. of having 20 heads = 0.0419.
- | $X \sim \text{Bin}(N, p)$, we have $E[X]=Np$, $\text{Var}(X)=Np(1-p)$

Confidence Interval for Accuracy

- | If N is large with high probability we get Np heads.
- | Then, if N is large enough we can easily estimate p (accuracy of our classifier) as $\text{\#heads}/N$.
- | N : how large? How confident on our estimation?
- | From *central limit theorem** it follows that we can approximate binomial distributions by normal distributions.....

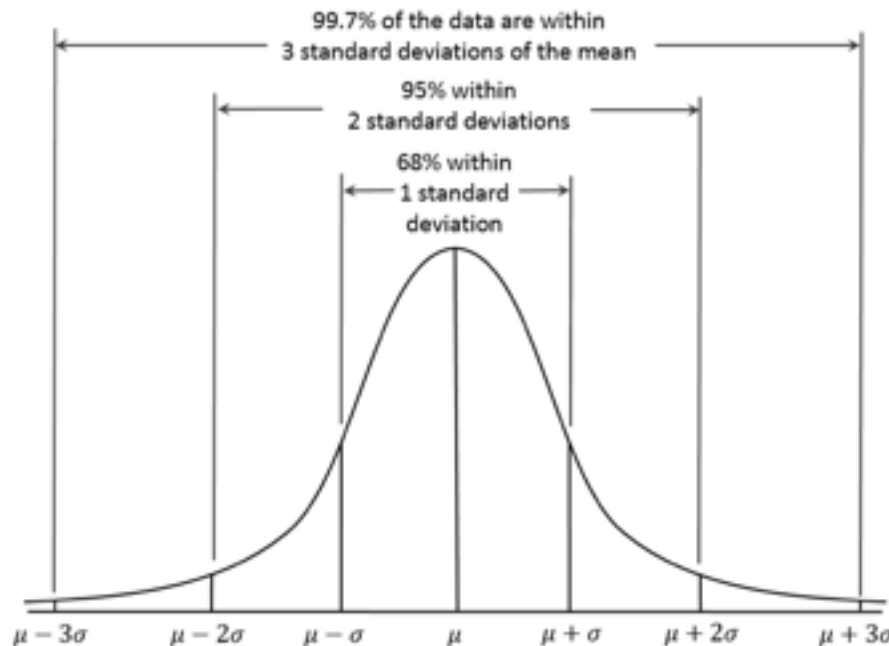
* arithmetic mean of a (sufficiently large) number of independent r.v. with variance σ^2 can be approximated by a normal distribution with variance σ^2 .

Normal distribution

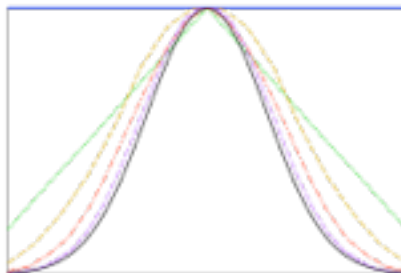
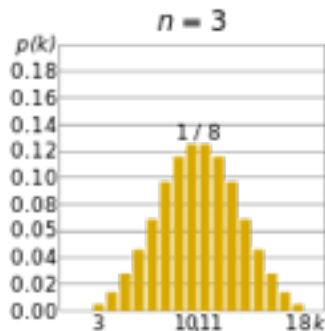
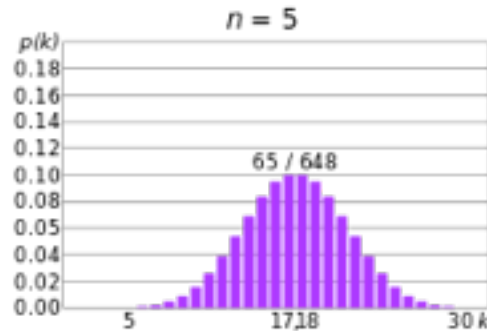
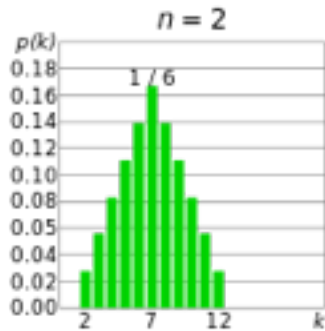
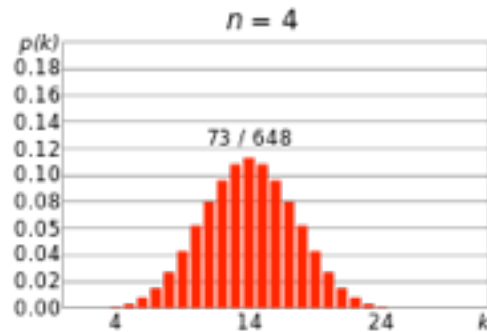
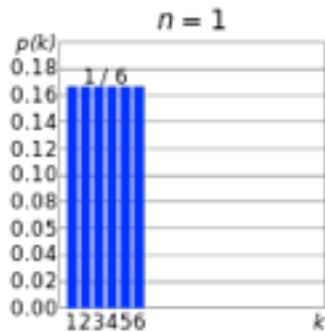
- Standard normal distribution $X \sim N(\mu, \sigma)$ is described by the following probability density function:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Intuitively, one can think of $P(x) dx$ as the probability of X falling within the infinitesimal interval $[x, x+dx]$



Approximating Binomial Distribution



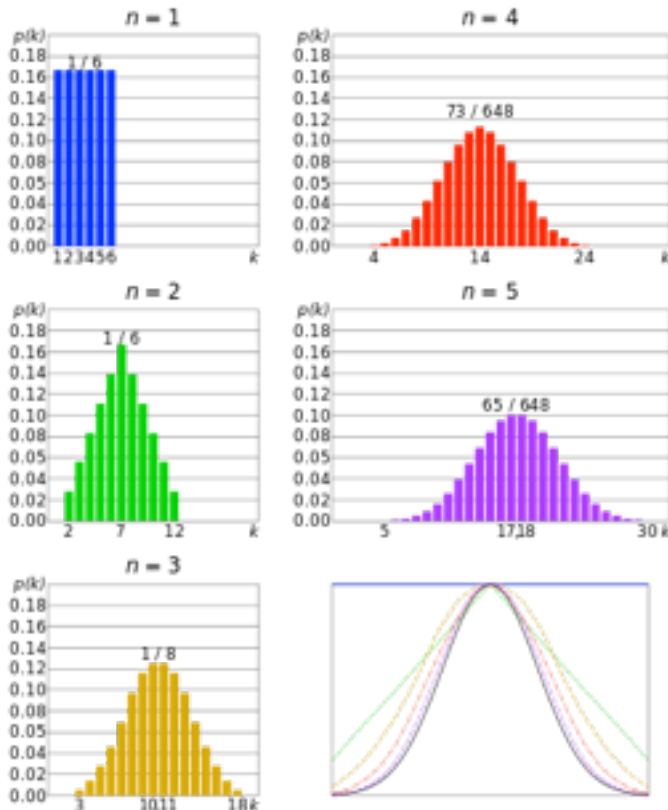
A dice is rolled n times, each graph shows the probability $p(k)$ that the sum of the values of the dice be k .

The last plot shows the normal distribution (black curve) and the other curves.

The normal distribution gives a good approximation as n gets larger.

Confidence Interval for Accuracy

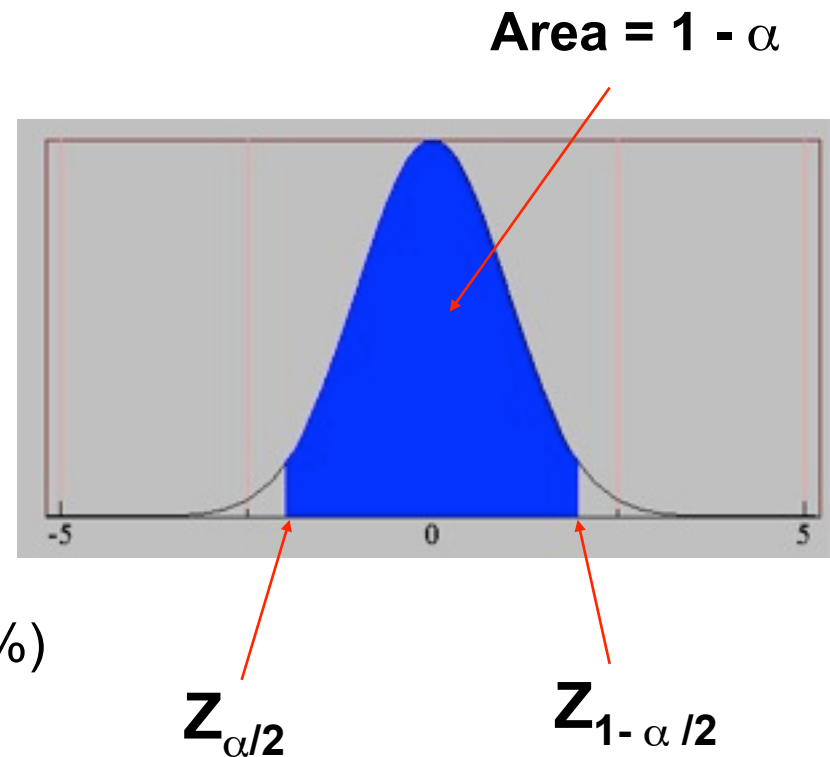
- What is the probability that the number of heads is in $[a,b]$? Approximated by the AUC (integral) of the normal distribution in $[a,b]$ (N large, e.g. $N > 30$)



Probability that the average sum of N dice rolls is in $[a,b]$ can be approximated by the integral in $[a,b]$ of the normal distribution (bottom right)

Confidence Interval for Accuracy

- For large test sets ($N > 30$), $\text{acc} = (\text{TP} + \text{TN}) / N$ has normal distribution with mean p and variance $\sigma^2 = p(1-p)/N$ ($N = \text{\#test instances}$, $p = \text{real accuracy of the classifier to be determined}$)
- Common values for the integrals of $Z \sim N(0,1)$ have been precomputed.
- Let Z_α be s.t. $P(Z \leq Z_\alpha) = \alpha$ ($\Rightarrow P(Z > 1 - Z_\alpha) = 1 - \alpha$).
- Focus on $1 - \alpha = 0.95$ (conf. 95%)



Confidence Interval for Accuracy

The following holds:

$$\begin{aligned} P(Z_{\alpha/2} < \frac{acc - p}{\sqrt{p(1-p)/N}} < Z_{1-\alpha/2}) \\ = 1 - \alpha \end{aligned}$$

From which we derive the CI for p with confidence $1 - \alpha$:

$$p = \frac{2 \times N \times acc + Z_{\alpha/2}^2 \pm \sqrt{Z_{\alpha/2}^2 + 4 \times N \times acc - 4 \times N \times acc^2}}{2(N + Z_{\alpha/2}^2)}$$

Confidence Interval for Accuracy

- | Consider a model that produces an accuracy of 80% when evaluated on 100 test instances:
 - $N=100$, $\text{acc} = 0.8$
 - Let $1-\alpha = 0.95$ (95% confidence)
 - From probability tables, $Z_{1-\alpha/2}=1.96$

N	50	100	500	1000	5000
p(lower)	0.670	0.711	0.763	0.774	0.789
p(upper)	0.888	0.866	0.833	0.824	0.811

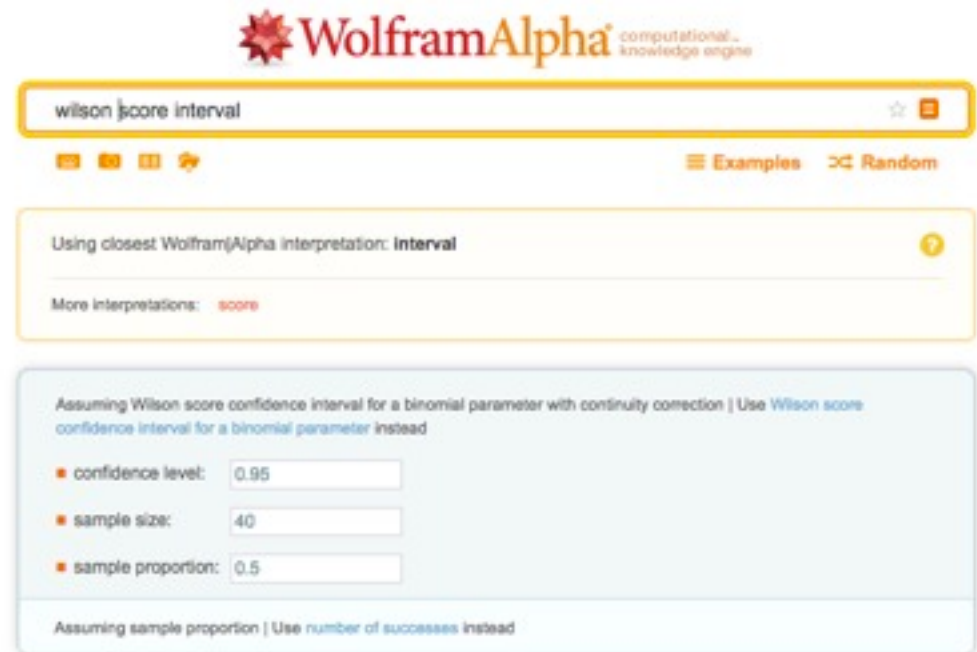
Intuition for Confidence Intervals

- | Prob. (in a sample of N test instances, Np test instances are classified successfully and p is in $[0.711, 0.866]$) = 0.95. In other words...
- | If we run the classifier again on a sample of N test instances, there is chance of 0.95 that Np test instances are classified correctly, p in $[0.711, 0.866]$.
- | As N gets larger, the confidence interval converges to the real accuracy of the classifier (i.e., when we run the classifier on all possible test instances).

Approximating Normal Distribution

- | Normal distribution:
 - does not work well if $N < 30$, or p is close to 0 or 1
 - rule of thumb: $Np > 5$ and $N(1-p) > 5$
- | More sophisticated techniques: Wilson score interval, Jeffrey intervals, Clopper-Pearson interval..

Most are available in
Wolfram Alpha



The screenshot shows the Wolfram Alpha interface. At the top, the search bar contains 'wilson score interval'. Below the search bar, there are icons for input methods (text, voice, image, etc.) and links for 'Examples' and 'Random'. The main content area displays the result: 'Using closest Wolfram|Alpha interpretation: Interval'. Below this, it says 'More interpretations: score'. At the bottom, there is a section titled 'Assuming Wilson score confidence interval for a binomial parameter with continuity correction | Use Wilson score confidence interval for a binomial parameter instead'. This section contains three input fields: 'confidence level: 0.95', 'sample size: 40', and 'sample proportion: 0.5'. At the very bottom, it says 'Assuming sample proportion | Use number of successes instead'.

A few additional comments on CI

- | We are making the assumption that test instances are drawn uniformly and independently at random, which is often not the case.
- | Observe that the results of the classifier on the test instances are usually not independent, this is not an issue.
- | Overall CI are reliable and give valuable information on the performance of a classifier.

Comparing Performance of 2 Models

- | Given two models, say M1 and M2, which is better?
 - M1 is tested on D1 (size= n_1), found error rate = e_1
 - M2 is tested on D2 (size= n_2), found error rate = e_2
 - Assume D1 and D2 are independent
 - If n_1 and n_2 are sufficiently large, then

$$e_1 \sim N(\mu_1, \sigma_1)$$

$$e_2 \sim N(\mu_2, \sigma_2)$$

- Approximate variance:

$$\hat{\sigma}_i^2 = \frac{e_i(1 - e_i)}{n_i}$$

Comparing Performance of 2 Models

- | To test if performance difference is statistically significant: $d = e1 - e2$
 - $d \sim N(d_t, \sigma_t)$ where d_t is the true difference
 - Since D1 and D2 are independent, their variance adds up:

$$\begin{aligned}\sigma_t^2 &= \sigma_1^2 + \sigma_2^2 \cong \hat{\sigma}_1^2 + \hat{\sigma}_2^2 \\ &= \frac{e1(1-e1)}{n1} + \frac{e2(1-e2)}{n2}\end{aligned}$$

- At $(1-\alpha)$ confidence level,

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

An Illustrative Example

- | Given: M1: $n1 = 30$, $e1 = 0.15$
M2: $n2 = 5000$, $e2 = 0.25$
- | $d = |e2 - e1| = 0.1$

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

- | At 95% confidence level, $Z_{\alpha/2} = 1.96$

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

=> **Interval contains 0** => difference not **statistically significant***.
With more advanced calculations: **good** with confidence $\leq 93.6\%$

* normal distribution does not approximate well binomial when p is small.