Freie Universität Berlin

Department Wirtschaftsinformatik

Fachbereich Wirtschaftswissenschaft

Business Intelligence

12 What is a good model?

Prof. Jan Fabian Ehmke, Sommersemester 2013

04.06.2013

Recommended reading

- Provost, F., Fawcett, T. — Data Science for Business, Chapter 7
- Berthold et al. — Guide to Intelligent Data Analysis, Chapter 5

Introduction

- What is desired from data mining results?

- How would you **measure** that your model is any good?
  - How to measure performance in a meaningful way?

- Model evaluation is **application-specific**
  - We look at common issues and themes in evaluation

- Frameworks and metrics for classification and instance scoring

Bad positives and harmless negatives

- Classification terminology
  - a **bad** outcome → a “positive” example [alarm!]
  - a **good** outcome → a “negative” example [uninteresting]
- Further examples
  - medical test: positive test → disease is present
  - fraud detector: positive test → unusual activity on account
- A classifier tries to distinguish the majority of cases (**negatives**, the uninteresting) from the small number of alarming cases (**positives**, alarming)
  - **number of mistakes** made on **negative** examples (false positive errors) will be relatively high
  - **cost of each mistake** made on a **positive** example (false negative error) will be relatively high

Agenda

- **Measuring accuracy**
  - Confusion matrix
  - Unbalanced classes
- A key analytical framework: Expected value
  - Evaluate classifier use
  - Frame classifier evaluation
- Evaluation and baseline performance

Measuring accuracy and its problems

- Up to now: measure a model's performance by some simple metric
  - classifier error rate, accuracy, …
- Simple example: accuracy

$$accuracy = \frac{\text{Number of correct decisions made}}{\text{Total number of decisions made}}$$

- Classification accuracy is popular, but usually **too simplistic** for applications of data mining to real business problems
- **Decompose** and count the different types of correct and incorrect decisions made by a classifier

The confusion matrix

- A confusion matrix for a problem involving n classes
  - is an $n \times n$ matrix with the columns labeled with actual classes and the rows labels with predicted classes

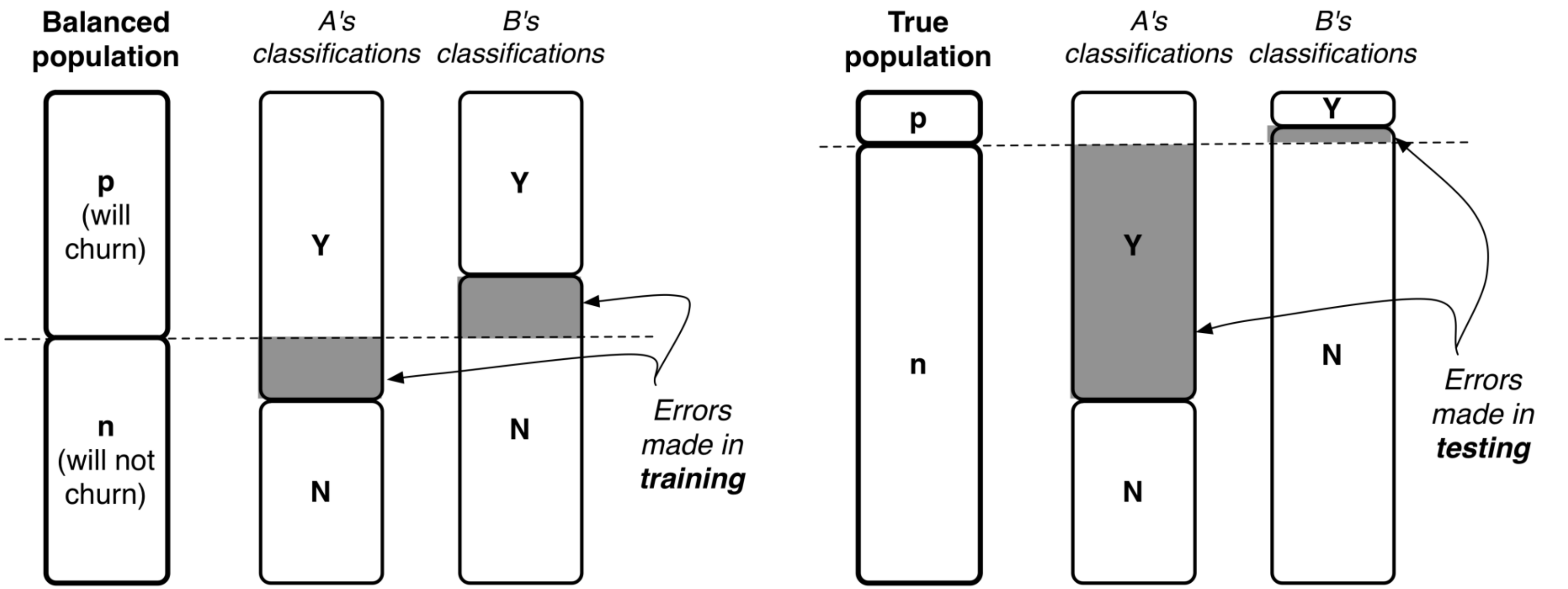
$$
\begin{array}{ccc}
 & & \begin{array}{cc} \quad\quad\quad\mathbf{p} \quad\quad\quad & \quad\quad\quad \mathbf{n} \quad\quad\quad \end{array} \\
\text{Predicted} & \begin{array}{c} \mathbf{Y} \\ \mathbf{N} \end{array} & \begin{pmatrix} \text{True positives} & \text{False positives} \\ \text{False negatives} & \text{True negatives} \end{pmatrix}
\end{array}
$$

- Each example in a test set has an **actual class label** and the **class predicted** by the classifier
- The confusion matrix separates out the decisions made by the classifier
  - actual/true classes: **p**(ositive), **n**(egative)
  - predicted classes: **Y**(es), **N**(o)
  - The main diagonal contains the count of correct decisions

Unbalanced classes (1/3)

- In practical classification problems, one class is often **rare**
  - Classification is used to find a relatively small number of **unusual ones** (defrauded customers, defective parts, targeting consumers who actually would respond, …)
  - The class distribution is unbalanced (“skewed”)
- Evaluation based on **accuracy** does not work
  - Example: 999:1 ratio – always choose the most prevalent class – 99.9% accuracy!
  - Fraud detection: skews of 10^2
  - Is a model with 80% accuracy always better than a model with 37% accuracy?
- We need to know more details about the population

Unbalanced classes (2/3)



- Consider two models A and B for the churn example (1000 customers, 1:9 ratio of churning)
  - Both models correctly classify 80% of the balanced pop.
  - Classifier A often falsely predicts that customers will churn
  - Classifier B makes many opposite errors

Unbalanced classes (3/3)

- Note the **different performances** of the models in form of a confusion matrix:

$$CM_A = \begin{array}{c c} & \begin{array}{cc} \textbf{churn} & \textbf{not churn} \end{array} \\ \begin{array}{c} Y \\ N \end{array} & \begin{pmatrix} 500 & 200 \\ 0 & 300 \end{pmatrix} \end{array}$$

$$CM_B = \begin{array}{c c} & \begin{array}{cc} \textbf{churn} & \textbf{not churn} \end{array} \\ \begin{array}{c} Y \\ N \end{array} & \begin{pmatrix} 300 & 0 \\ 200 & 500 \end{pmatrix} \end{array}$$

- Model A achieves 80% accuracy on the balanced sample
- Unbalanced population: A‘s accuracy is 37%, B‘s accuracy is 93%
- Which model is better?

Unequal costs and benefits

- How much do we care about the different **errors** and correct decisions?
  - Classification accuracy makes no distinction between **false positive** and **false negative** errors
  - In real-world applications, different kinds of errors lead to different consequences!
- Examples for medical diagnosis:
  - a patient has cancer (although he does not)
    → **false positive error**, expensive, but not life threatening
  - a patient has cancer, but she is told that she has not
    → **false negative error**, more serious
- Errors should be counted separately
  - Estimate cost or benefit of each decision

A look beyond classification

- Another example: how to measure the accuracy / quality of a regression model?
  - Predict how much a given customer will like a given movie

- Typical accuracy of regression: mean-squared error

- What does the mean-squared error describe?
  - Value of the target variable, e.g., the number of stars that a user would give as a rating for the movie

- Is the mean-squared error a meaningful metric?

Agenda

- Measuring accuracy
  - Confusion matrix
  - Unbalanced classes
- **A key analytical framework: Expected value**
  - Evaluate classifier use
  - Frame classifier evaluation
- Evaluation and baseline performance

The expected value framework

- Expected value calculation includes **enumeration of the possible outcomes** of a situation
- Expected value = weighted average of the values of different possible outcomes, where the weight given to each value is the probability of its occurrence
  - Example: different levels of profit
  - We focus on the maximization of expected profit
- **General form** of expected value computation:

$$EV = p(o_1) \cdot v(o_1) + p(o_2) \cdot v(o_2) + \cdots +$$

with o_i as possible decision outcome,
$p(o_i)$ as its probability, and $v(o_i)$ as its value.

- Probabilities can be **estimated** from available data

Expected value for use of a classifier (1/2)

- **Use of a classifier:** predict a class and take some action
  - Example target marketing: assign each consumer to either a class „likely responder“ or „not likely responder“
  - Response is usually relatively low – so no consumer may seem like a likely responder
- Computation of the expected value
  - A model gives an estimated probability of response $\hat{p}_R(\boldsymbol{x})$ for any consumer with a feature vector $\boldsymbol{x}$
  - **Calculate expected benefit (or costs)** of targeting consumer $\boldsymbol{x}$: $\hat{p}_R(\boldsymbol{x}) \cdot v_R + (1 - \hat{p}_R(\boldsymbol{x})) \cdot v_{NR}$
    with v_R being the value of a response and
    v_{NR} the value from no response

Expected value for use of a classifier (2/2)

- Example
  - Price of product: \$200, costs of product: \$100
  - Targeting a consumer: \$1, profit $v_R = \$99$, $v_{NR} = -\$1$
  - Do we make a profit? Is the expected value (profit) of targeting greater than zero?

$$\hat{p}_R(\boldsymbol{x}) \cdot \$99 + (1 - \hat{p}_R(\boldsymbol{x})) \cdot (-\$1) > 0$$
$$\hat{p}_R(\boldsymbol{x}) \cdot \$99 > (1 - \hat{p}_R(\boldsymbol{x})) \cdot \$1$$
$$\hat{p}_R(\boldsymbol{x}) > 0.01$$

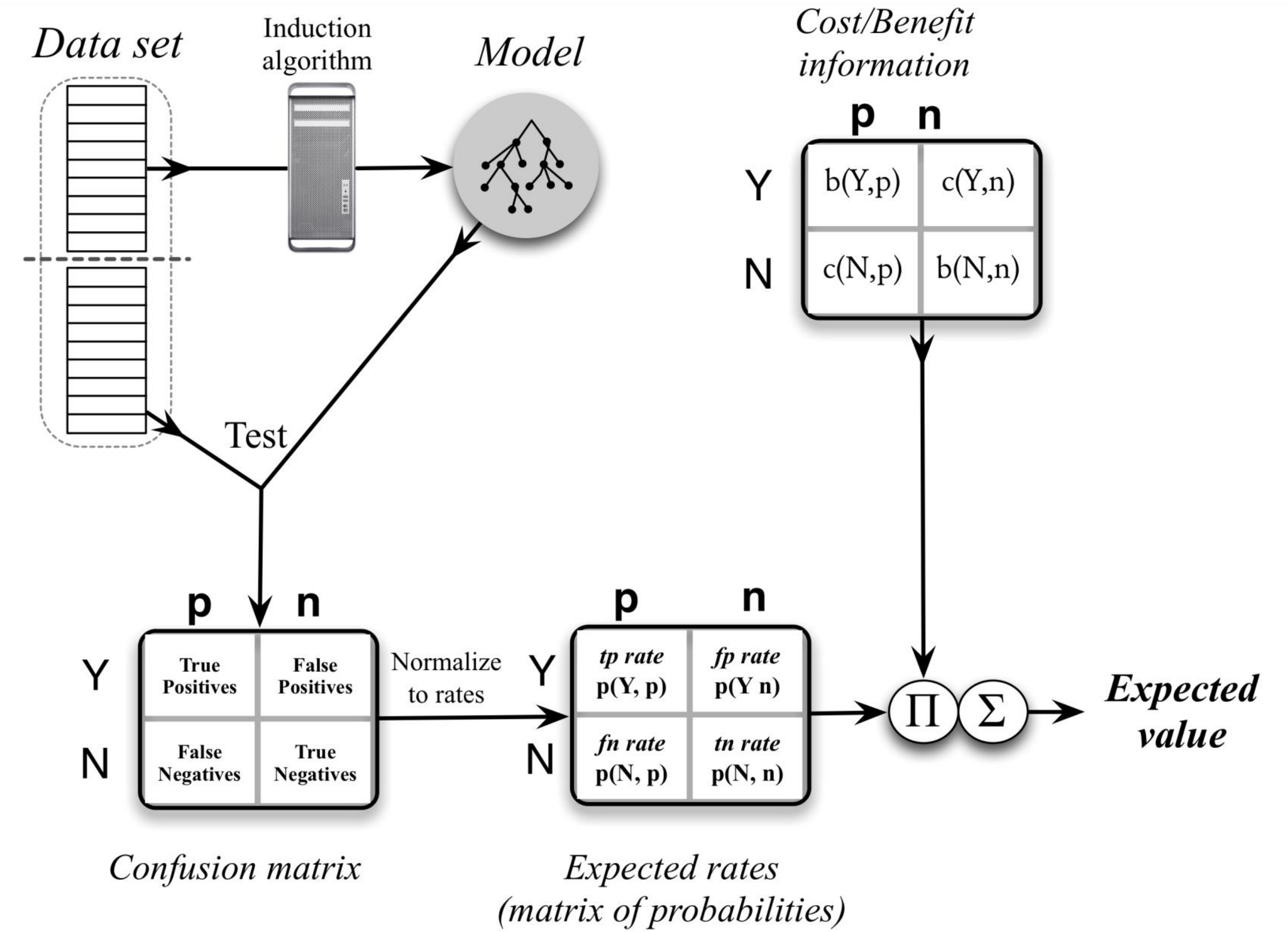
  - We should target the consumer as long as the estimated probability of responding is greater than 1%!

Expected value for evaluation of a classifier

- Goal: **compare the quality of different models** with each other
  - Does the data-driven model perform better than a hand-crafted model?
  - Does a classification tree work better than a linear discriminant model?
  - Do any of the models perform substantially better than a baseline model?

- **In aggregate:** how well does each model do – what is its expected value?

Expected value calculation



Expected value for evaluation of a classifier

- **Aggregate** together all the different cases:
  - When we target consumers, what is the probability that they (do not) respond?
  - What about when we do not target consumers, would they have responded?
- This information is available in the confusion matrix
  - Each o_i corresponds to one of the possible combinations of the class we predict/the actual class
- Example confusion matrix/estimates of probability

Predicted \ Actual	p	n
Y	56	7
N	5	42

$T = 100, P = 61, N = 49$ (**P**ositive, **N**egative)

$$p(Y, \mathrm{p}) = \frac{56}{100} = 0.56, p(Y, \mathrm{n}) = \frac{7}{100} = 0.7$$

$$p(N, \mathrm{p}) = \frac{5}{100} = 0.05, p(N, \mathrm{n}) = \frac{42}{100} = 0.42$$

Error rates

- Where do the probabilities of errors and correct decisions actually come from?

- Each cell of the confusion matrix contains a count of the number of decisions corresponding to the combination of (predicted, actual) $count(h, a)$

- Compute estimated probabilities as
$$p(h, a) = count(h, a)/T$$

Costs and benefits

- Compute **cost-benefit values** for each decision pair
- A cost-benefit matrix specifies for each (predicted,actual) pair the cost or benefit making such a decision
  - Correct classifications (true positives and negatives) correspond to $b(Y,p)$ and $b(N,n)$, respectively
  - Incorrect classifications (false positives and negatives) correspond to $b(Y,n)$ and $b(N,n)$, respectively [often negative benefits or costs]

		Actual	
		p	**n**
Predicted	**Y**	b(**Y,p**)	c(**Y,n**)
	N	c(**N,p**)	b(**N,n**)

- Costs and benefits cannot be estimated from data
  - How much is it really worth us to retain a customer?
  - Often use of average estimated costs and benefits

Costs and benefits - example

- Targeted marketing example
  - **False positive** occurs when we classify a consumer as a likely responder and therefore target her, but she does not respond → benefit $b(Y, \mathrm{n}) = -1$
  - **False negative** is a consumer who was predicted not to be a likely responder, but would have bought if offered. No money spent, nothing gained → benefit $b(N, \mathrm{p}) = 0$
  - **True positive** is a consumer who is offered the product and buys it → benefit $b(Y, \mathrm{p}) = 200 - 100 - 1 = 99$
  - **True negative** is a consumer who was not offered a deal but who would not have bought it → benefit $b(N, \mathrm{n}) = 0$
- Sum up in cost-benefit matrix

		Actual	
Predicted		p	n
	Y	99	-1
	N	0	0

Expected profit computation (1/2)

- Compute **expected profit** by cell-wise multiplication of the matrix of costs and benefits against the matrix of probabilities:

$$EP = p(Y|\mathrm{p}) \cdot p(\mathrm{p}) \cdot b(Y,\mathrm{p}) + p(N|\mathrm{p}) \cdot p(\mathrm{p}) \cdot b(N,\mathrm{p}) + p(N|\mathrm{n}) \cdot p(\mathrm{n}) \cdot b(N,\mathrm{n}) + p(Y|\mathrm{n}) \cdot p(\mathrm{n}) \cdot b(Y,\mathrm{n})$$

- Sufficient for comparison of various models

- Alternative calculation: **factor out the probabilities** of seeing each class (class priors)
  - Class priors $p(\mathrm{p})$ and $p(\mathrm{n})$ specify the likelihood of seeing positive versus negative instances
  - Factoring out allows us to separate the influence of class imbalance from the predictive power of the model

Expected profit computation (2/2)

- Factoring out priors yields the following **alternative expression** for expected profit

$$EP = p(\mathrm{p}) \cdot [p(Y|\mathrm{p}) \cdot b(Y,\mathrm{p}) + p(N|\mathrm{p}) \cdot b(N,\mathrm{p})] + p(\mathrm{n}) \cdot [p(N|\mathrm{n}) \cdot b(N,\mathrm{n}) + p(Y|\mathrm{n}) \cdot b(Y,\mathrm{n})]$$

- The first component corresponds to the expected profit from the **positive examples**, whereas the second corresonds to the expected profit from the **negative examples**

Costs and benefits – example alternative expression

		Actual p	Actual n
Predicted	**Y**	56	7
	N	5	42

$$P = 61, \qquad N = 49$$
$$p(\mathbf{p}) = 0.55, \qquad p(\mathbf{n}) = 0.45$$
$$tp\ rate = 56/61 = 0.92, \qquad fp\ rate = 7/49 = 0.14$$
$$fn\ rate = 5/61 = 0.08, \qquad tn\ rate = 42/49 = 0.86$$

$$\begin{aligned} expected\ profit &= p(\mathbf{p}) \cdot [p(\mathbf{Y}|\mathbf{p}) \cdot b(\mathbf{Y},\mathbf{p}) + p(\mathbf{N}|\mathbf{p}) \cdot b(\mathbf{N},\mathbf{p})] + \\ &\quad p(\mathbf{n}) \cdot [p(\mathbf{N}|\mathbf{n}) \cdot b(\mathbf{N},\mathbf{n}) + p(\mathbf{Y}|\mathbf{p}) \cdot b(\mathbf{Y},\mathbf{n})] \\ &= 0.55 \cdot [0.92 \cdot b(\mathbf{Y},\mathbf{p}) + 0.08 \cdot b(\mathbf{N},\mathbf{p})] + \\ &\quad 0.45 \cdot [0.86 \cdot b(\mathbf{N},\mathbf{n}) + 0.14 \cdot b(\mathbf{Y},\mathbf{n})] \\ &= 0.55 \cdot [0.92 \cdot 100 + 0.08 \cdot 0] + \\ &\quad 0.45 \cdot [0.86 \cdot 0 + 0.14 \cdot -1] \\ &= 50.6 - 0.063 \approx \mathbf{50.54} \end{aligned}$$

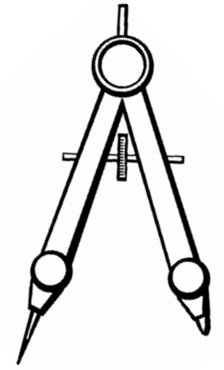
- This expected value means that if we apply this model to a population of prospective customers and mail offers to those it classifies as positive, we can expect to make an average of about $50.54 profit per consumer.

Further insights

- In sum: instead of computing accuracies for competing models, we would compute expected values
- We can compare two models even though one is based on a representative distribution and one is based on a class-balanced data set
  - Just replace the priors
  - Balanced distribution → $p(\mathbf{p}) = 0.5$ and $p(\mathbf{n}) = 0.5$
- Make sure that the signs of quantities in the cost-benefit matrix are consistent
- Do not double count by putting a benefit in one cell and a negative cost for the same thing in another cell

Other evaluation metrics

- Based on the entries of the confusion matrix, we can describe various evaluation metrics
  - True positive rate (Recall): $\frac{TP}{TP+FN}$
  - False negative rate: $\frac{FN}{TP+FN}$
  - Precision (accuracy over the cases predicted to be positive): $\frac{TP}{TP+FP}$
  - F-measure (harmonic mean): $2 \cdot \frac{precision \cdot recall}{precision+recall}$
  - Sensitivity: $\frac{TN}{TN+FP}$
  - Specificity: $\frac{TP}{TP+FN}$
  - Accuracy (count of correct decisions): $\frac{TP+TN}{P+N}$



Agenda

- Measuring accuracy
  - Confusion matrix
  - Unbalanced classes
- A key analytical framework: Expected value
  - Evaluate classifier use
  - Frame classifier evaluation
- **Evaluation and baseline performance**

Baseline performance (1/3)

- Consider what would be a reasonable baseline against which to compare model performance
  - Demonstrate stakeholder that data mining has added value (or not)
- What is the appropriate baseline for comparison?
  - Depends on the actual application
- Nate Silver on weather forecasting:
  - *There are two basic tests that any weather forecast must pass to demonstrate its merit: (1) It must do better than what meteorologists call persistence: the assumption that the weather will be the same tomorrow (and the next day) as it was today. (2) It must also beat climatology, the long-term historical average of conditions on a particular date in a particular area.*

Baseline performance (2/3)

- Baseline performance for classification
  - Compare to a completely random model (very easy)
  - Implement a simple (but not simplistic) alternative model
- **Majority classifier** = a naive classifier that always chooses the majority class of the training data set
  - May be challenging to outperform: classification accuracy of 94%, but only 6% of the instances are positive
    → majority classifier also would have an accuracy of 94%!
- Pitfall: don‘t be surprised that many models simply predict everything to be of the majority class
- Maximizing simple prediction accuracy is usually not an appropriate goal

Baseline performance (3/3)

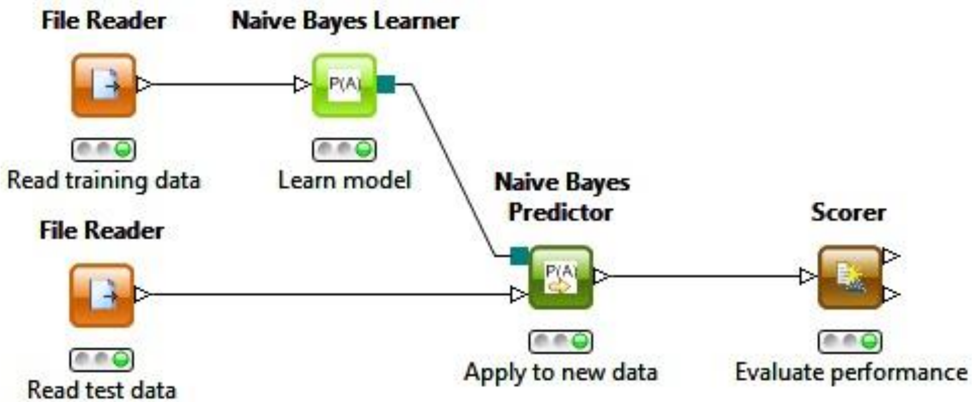
- Further alternative: how well does a simple “conditional” model perform?
  - Conditional → prediction different based on the value of the features
  - Just use the most informative variable for prediction
  - Decision tree: build a tree with only one internal node (decision stump) → tree induction selects the single most informative feature to make a decision
- Compare quality of models based on data sources
  - Quantify the value of each source
- Implement models that are based on domain knowledge

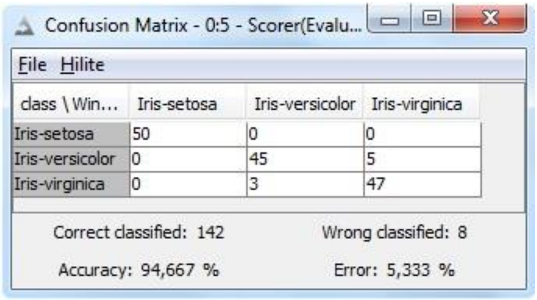
Agenda

- Measuring accuracy
  - Confusion matrix
  - Unbalanced classes
- A key analytical framework: Expected value
  - Evaluate classifier use
  - Frame classifier evaluation
- Evaluation and baseline performance

Example with KNIME

- The scorer node is KNIME's most prominent module to estimate errors.
  - In the figure below, the trained Naïve Bayes classifier is applied to a second data set, and the output is fed into the scorer node which compares the target with the predicted class.
  - The output of this scorer is a confusion matrix and a second matrix listing some well-known error measures.



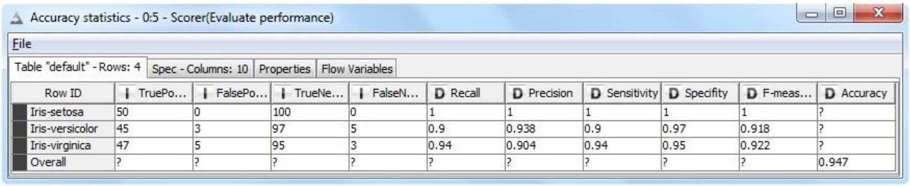

Confusion Matrix - 0:5 - Scorer(Evalu...

File Hilite

class \ Win...	Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	50	0	0
Iris-versicolor	0	45	5
Iris-virginica	0	3	47

Correct classified: 142 Wrong classified: 8

Accuracy: 94,667 % Error: 5,333 %


Accuracy statistics - 0:5 - Scorer(Evaluate performance)

File

Table "default" - Rows: 4 | Spec - Columns: 10 | Properties | Flow Variables

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specifity	F-meas...	Accuracy
Iris-setosa	50	0	100	0	1	1	1	1	1	?
Iris-versicolor	45	3	97	5	0.9	0.938	0.9	0.97	0.918	?
Iris-virginica	47	5	95	3	0.94	0.904	0.94	0.95	0.922	?
Overall	?	?	?	?	?	?	?	?	?	0.947