Chapter 5 – Evaluating Classification & Predictive Performance

Instructor: Zach Zhizhong ZHOU, Shanghai Jiao Tong University
主讲教师：周志中，上海交通大学

Data Mining for Business Intelligence

Shmueli, Patel & Bruce

Why Evaluate? 为何要评估结果？

- Multiple methods are available to classify or predict 有多种方法可用于分类或预测。

- For each method, multiple choices are available for settings 对每一种方法，有多个模型设置的选择。

- To choose best model, need to assess each model’s performance 为了找到最佳模型，需要评估每个模型的表现情况。

Accuracy Measures (Classification)
准确性度量（分类）

Misclassification error 误分类错误

- Error = classifying a record as belonging to one class when it belongs to another class. 错误 = 将本属于一类的记录划归另外一类。

- Error rate = percent of misclassified records out of the total records in the validation data 错误率：误分类记录在测试数据集中所占比率。

Naïve Rule 朴素分类准则

Naïve rule: classify all records as belonging to the most prevalent class 朴素分类准则：将所有记录划为最普遍的一类。

- Often used as benchmark: we hope to do better than that 经常作为基准：我们希望比它做得更好。
- Exception: when goal is to identify high-value but rare outcomes, we may do well by doing worse than the naïve rule (see “lift” – later) 例外：当目标是找到高价值但少见的结果，那么即使分类结果比朴素分类准则的准确率更差，我们也可以得到一个好的分类结果。

Separation of Records 记录分离

“High separation of records” means that using predictor variables attains low error “高效记录分离”意思是使用预测因子获得低错误率。

“Low separation of records” means that using predictor variables does not improve much on naïve rule “低效记录分离”意味着使用预测因子未能提高朴素分类准则的预测结果。

Confusion Matrix 混淆矩阵

201 1’s correctly classified as “1”

85 1’s incorrectly classified as “0”

25 0’s incorrectly classified as “1”

2689 0’s correctly classified as “0”

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	201	85
0	25	2689

Error Rate 错误率

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	201	85
0	25	2689

Overall error rate = (25+85)/3000 = 3.67%

Accuracy = 1 – err = (201+2689) = 96.33%

If multiple classes, error rate is:

(sum of misclassified records)/(total records)

Cutoff for classification 分类的截值

Most DM algorithms classify via a 2-step process: 绝大多数数据挖掘算法通过2个步骤进行分类：

For each record, 对每一个记录

1. Compute probability of belonging to class “1” 计算该记录属于类别1的概率 （假设只有2个类别）
2. Compare to cutoff value, and classify accordingly 与截值比较，然后决定是否将该记录归入类别1。

- Default cutoff value is 0.50 截值默认为0.5

  If >= 0.50, classify as “1” 如大于0.5，类别视为1

  If < 0.50, classify as “0” 否则视为0

- Can use different cutoff values 可以使用不同的截值
- Typically, error rate is lowest for cutoff = 0.50 特别地，错误率在截值为0.5时达到最低。

Cutoff Table 截值分类表

Actual Class	Prob. of "1"	Actual Class	Prob. of "1"
1	0.996	1	0.506
1	0.988	0	0.471
1	0.984	0	0.337
1	0.980	1	0.218
1	0.948	0	0.199
1	0.889	0	0.149
1	0.848	0	0.048
0	0.762	0	0.038
1	0.707	0	0.025
1	0.681	0	0.022
1	0.656	0	0.016
0	0.622	0	0.004

- If cutoff is 0.50: eleven records are classified as “1”
- If cutoff is 0.80: seven records are classified as “1”

Confusion Matrix for Different Cutoffs

Cut off Prob.Val. for Success (Updatable)	**0.25**

Classification Confusion Matrix		
	Predicted Class	
Actual Class	owner	non-owner
owner	11	1
non-owner	4	8

Cut off Prob.Val. for Success (Updatable)	**0.75**

Classification Confusion Matrix		
	Predicted Class	
Actual Class	owner	non-owner
owner	7	5
non-owner	1	11

Lift 提升水平

When One Class is More Important 当某种类别更为重要

In many cases it is more important to identify members of one class 很多情况下识别某种类别的成员更为重要

- Tax fraud 税务欺诈
- Credit default 信用破产
- Response to promotional offer 对促销优惠的应答
- Detecting electronic network intrusion 检测识别网络入侵
- Predicting delayed flights 预测航班延误

In such cases, we are willing to tolerate greater overall error, in return for better identifying the important class for further attention 在这些情况下，我们愿意容忍高一点的错误率，以便更好识别重要类别的成员。

Alternate Accuracy Measures 替代的准确率度量指标

If “C_1” is the important class, 如果C1是重要的类别。

Sensitivity = % of “C_1” class correctly classified
敏感度 = 被正确判定为C1占C1总记录的比率

Specificity = % of “C_0” class correctly classified
特异度=被正确判定为C0占C0中记录的比率的比率

False positive rate = % of predicted “C_1’s” that were not “C_1’s” 假阳性率：被错误判定为C1的记录占C0总记录的比率。

False negative rate = % of predicted “C_0’s” that were not “C_0’s” 假阴性率：被错误判定为C0的记录占C1总记录的比率。

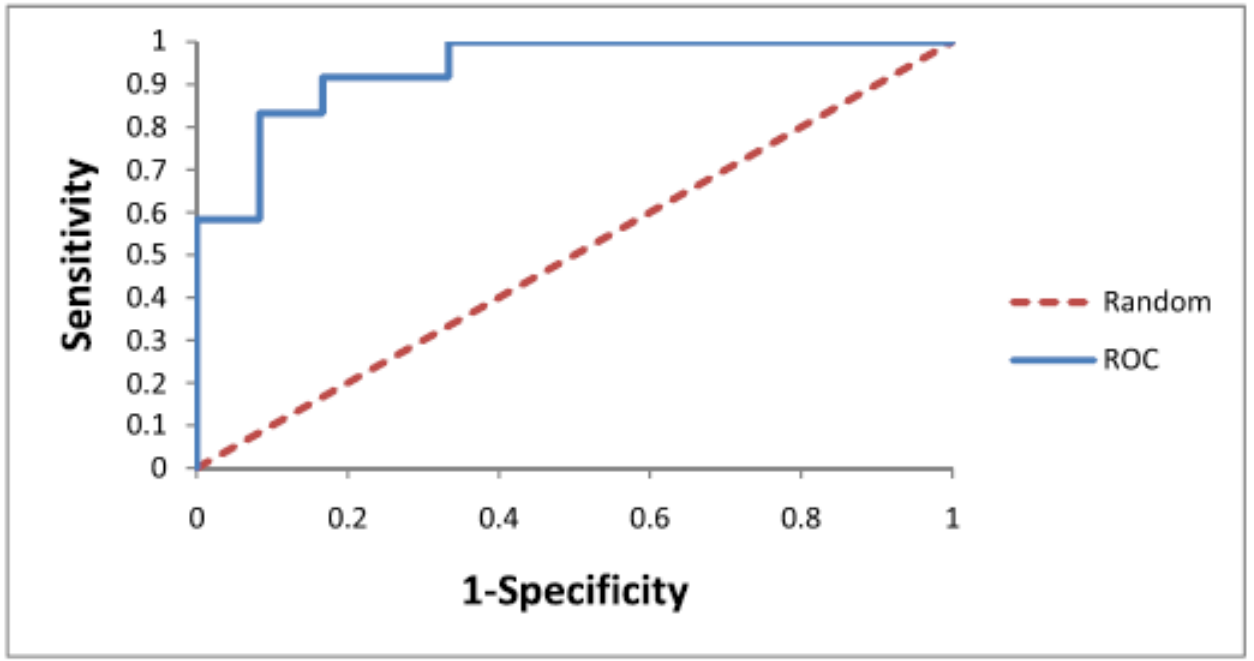
Alternate Accuracy Measures 替代的准确率度量指标

true positive (TP) : with hit
true negative (TN): with correct rejection
false positive (FP): with false alarm, Type I error
false negative (FN): with miss, Type II error

Sensitivity = TP / P = TP / (TP + FN)
Specificity = TN / N = TN / (TN + FP)
False Positive Rate = FP / N = FP / (TN+FP)
False Negative Rate = FN / P = FN / (TP + FN)

ROC Curve (Receiver Operating Characteristic Curve 接收者操作特征曲线)

被正确判定为C1的记录占C1总记录的比率。



被错误判定为C1的记录占C0总记录的比率。

Lift and Decile Charts: Goal 提升水平和十分位数表：目标

Useful for assessing performance in terms of identifying the most important class 可用于评估模型识别重要类别成员的表现

Helps evaluate, e.g., 有助于估计

- How many tax records to examine 多少税表需要检查？
- How many loans to grant 应该批准多少额度的贷款？
- How many customers to mail offer to 应该给多少消费者寄送优惠券？

Lift and Decile Charts – Cont. 提升水平和十分位数表

Compare performance of DM model to “no model, pick randomly” 将模型与“没有模型，随机挑选”进行比较

Measures ability of DM model to identify the important class, relative to its average prevalence 衡量数据挖掘模型相对于重要类别成员出现率的识别重要类别的能力。

Charts give explicit assessment of results over a large number of cutoffs 图表给出对于大量的截值的清楚的估计结果。

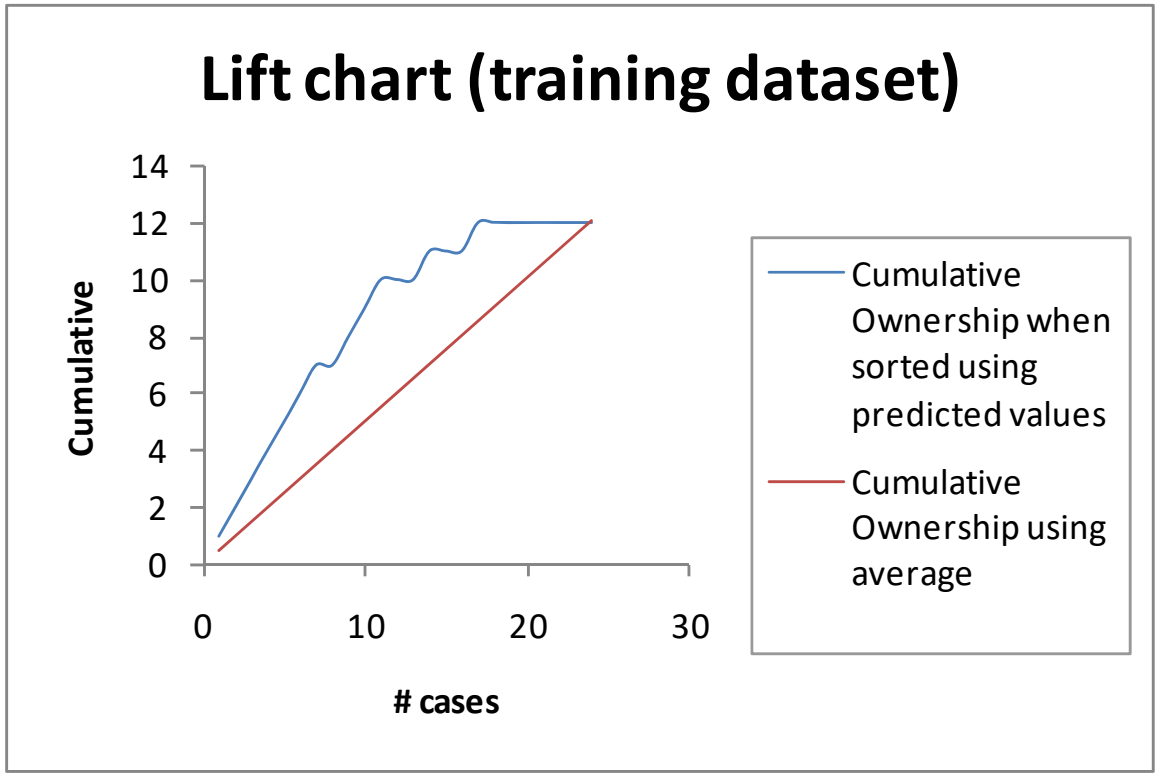
Lift and Decile Charts: How to Use
提升水平和十分位数表：如何使用

Compare lift to “no model” baseline 与无模型的基准进行比较

In lift chart: compare step function to straight line 在提升水平图中比较阶跃函数与直线。

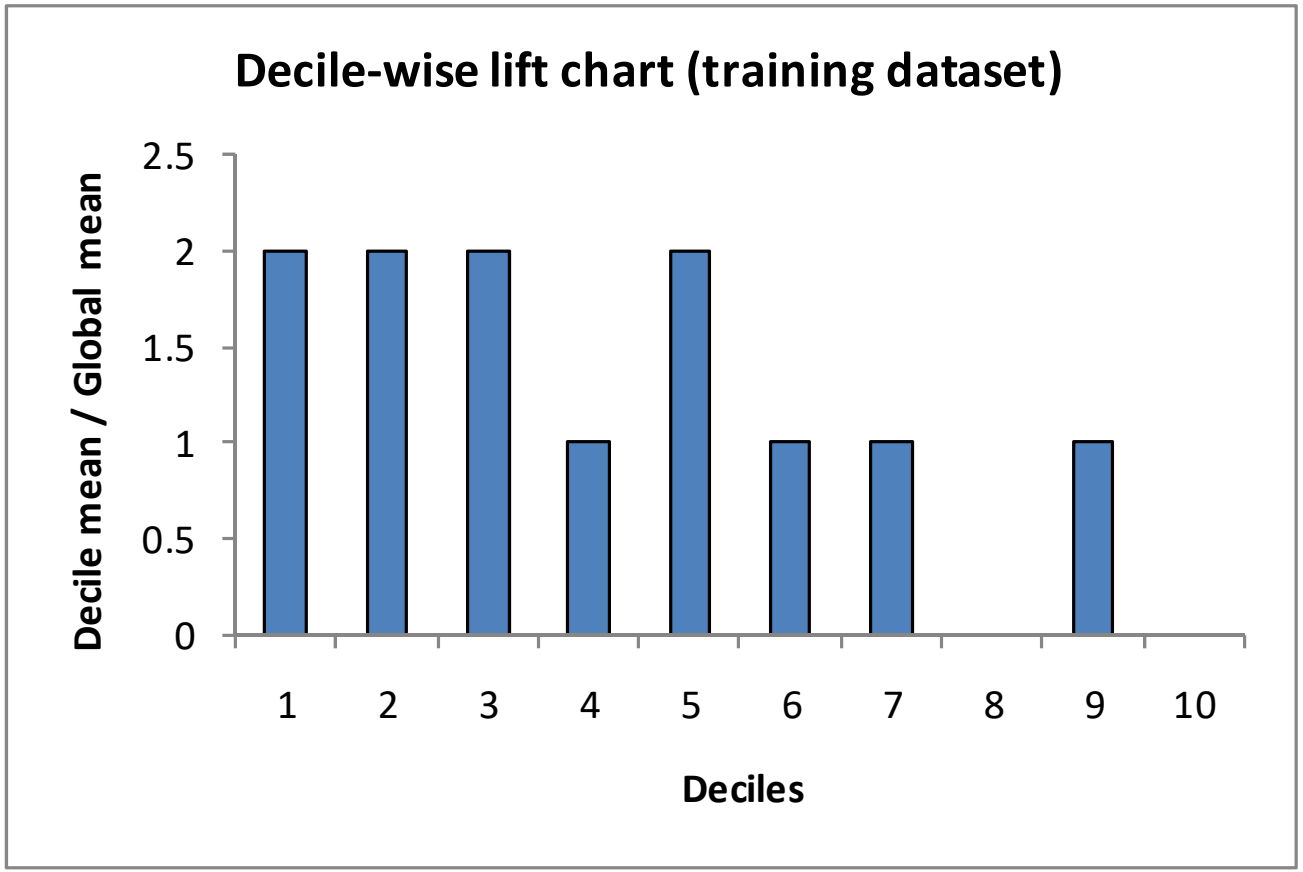
In decile chart compare to ratio of 1 在十分位数表中与1进行比较。

Lift Chart – cumulative performance
提升水平图：累积表现



After examining (e.g.,) 10 cases (x-axis), 9 owners (y-axis) have been correctly identified

Decile Chart 十分位数表



In "most probable" (top) decile, model is twice as likely to identify the important class (compared to avg. prevalence)

Lift Charts: How to Compute 如何计算提升水平表？

- Using the model’s classifications, sort records from most likely to least likely members of the important class 使用模型算出归入重要类别的可能性按取值从大到小排序。

- Compute lift: Accumulate the correctly classified “important class” records (Y axis) and compare to number of total records (X axis) 对于记录总量X，算出有多少记录Y正确地被归类为重要类别。

Lift vs. Decile Charts 提升水平和十分位数表比较

Both embody concept of “moving down” through the records, starting with the most probable 两者都使用了从最可能归为重要类别的记录往下移动的概念。

Decile chart does this in decile chunks of data

- Y axis shows ratio of decile mean to overall mean

十分位数表评估在离散的十分位数上的Y的取值。

Lift chart shows continuous cumulative results

- Y axis shows number of important class records identified

□而提升水平图评估的是连续的X上Y的取值。

Asymmetric Costs 不对称成本

Misclassification Costs May Differ
错误分类的成本可能有所不同

The cost of making a misclassification error may be higher for one class than the other(s) 对某个类别的错误分类成本可能高于对另外一个类别的错误分类成本。

Looked at another way, the benefit of making a correct classification may be higher for one class than the other(s) 对一个类别的正确分类带来的效益可能高于对于另外一个类别的正确分类所带来的效益。

Example – Response to Promotional Offer
例子：销售推广优惠的应答率

Suppose we send an offer to 1000 people, with 1% average response rate 优惠券发给100人只有1%的平均应答率

(“1” = response, “0” = nonresponse)

- “Naïve rule” (classify everyone as “0”) has error rate of 1% (seems good)

- Using DM we can correctly classify eight 1’s as 1’s

  It comes at the cost of misclassifying twenty 0’s as 1’s and two 0’s as 1’s.

The Confusion Matrix

	Predict as 1	Predict as 0
Actual 1	8	2
Actual 0	20	970

Error rate = (2+20) = 2.2% (higher than naïve rate)

Introducing Costs & Benefits

Suppose:

- Profit from a "1" is $10
- Cost of sending offer is $1

Then:

- Under naïve rule, all are classified as "0", so no offers are sent: no cost, no profit
- Under DM predictions, 28 offers are sent.
  - 8 respond with profit of $10 each
  - 20 fail to respond, cost $1 each
  - 972 receive nothing (no cost, no profit)
- Net profit = $60

Profit Matrix

	Predict as 1	Predict as 0
Actual 1	$80	0
Actual 0	($20)	0

Lift (again)

Adding costs to the mix, as above, does not change the actual classifications 对混淆矩阵的成本效益分析不改变分类结果。

Better: Use the lift curve and change the cutoff value for “1” to maximize profit 最好使用提升水平曲线但把截值从1改成最大化的利润。

Generalize to Cost Ratio 推广到成本比率

Sometimes actual costs and benefits are hard to estimate 有时实际的成本和收益难以估计。但我们……

- Need to express everything in terms of costs (i.e., cost of misclassification per record) 需要把所有指标用成本表达出来（例如，因错误分类导致的每个记录的平均成本）

- Goal is to minimize the average cost per record 目标是降低单位记录的平均成本。

A good practical substitute for individual costs is the ratio of misclassification costs (e,g,, “misclassifying fraudulent firms is 5 times worse than misclassifying solvent firms”) 一个现实的替代方案是用成本比率换掉错误分类造成的单次成本。（比如，对骗子企业的错误分类成本5倍于对健康企业的错误分类）

Minimizing Average Misclassification Cost
降低平均错分类成本

q_1 = cost of misclassifying an actual "1",

q_0 = cost of misclassifying an actual "0"

$$\left(q_0 n_{0,1} + q_1 n_{1,0}\right)/n = q_0\left[n_{0,1} + \left(q_1/q_0\right)n_{1,0}\right]/n$$

The minimization only depends on q1/q0

对该式子进行最小化只需要知道成本比率就可以了。

Software may provide option for user to specify cost ratio

Multiple Classes 多个分类

For *m* classes, confusion matrix has *m* rows and *m* columns 对于m类，混淆矩阵有m行m列。

- Theoretically, there are *m*(*m*-1) misclassification costs, since any case could be misclassified in *m*-1 ways 理论上有m(m-1)个误分类成本。因为每个case有m-1种误分类。
- Practically too many to work with 实际操作中有太多计算。
- In decision-making context, though, such complexity rarely arises – one class is usually of primary interest 但是实际决策中这种复杂度很少成为问题，因为通常只有一个分类是重点关注的。

Adding Cost/Benefit to Lift Curve
在提升曲线上加上成本收益分析

- Sort records in descending probability of success 将记录按照成功分入该类的概率从高到低进行排序。

- For each case, record cost/benefit of actual outcome 对每个案子，记录成本收益分析的实际结果。

- Also record cumulative cost/benefit 同时记录累积成本和收益。

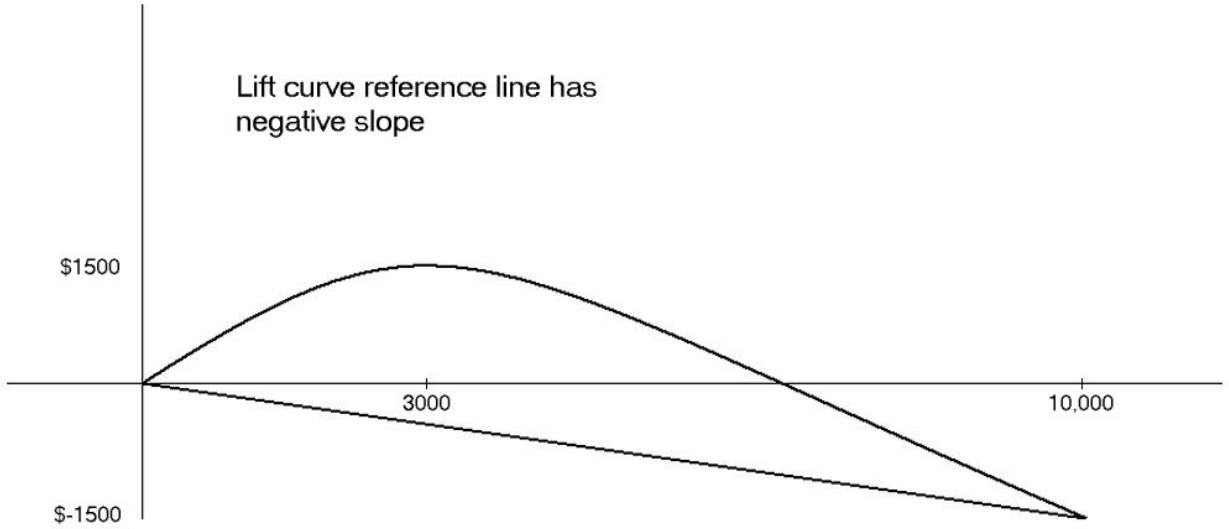
- Plot all records 将所有记录在图上表达出来。
  - X-axis is index number (1 for 1st case, n for nth case) x轴代表从1到n的记录。
  - Y-axis is cumulative cost/benefit y轴代表累积成本或者收益。
  - Reference line from origin to y_n (y_n = total net benefit) 参照直线连接原点和y_n (y_n 代表总净收益)

Lift Curve May Go Negative 提升曲线可以是负斜率

If total net benefit from all cases is negative, reference line will have negative slope 如果从所有案子获得的净收益是负数，那么参照直线是负斜率。

Nonetheless, goal is still to use cutoff to select the point where net benefit is at a maximum 但是，目标仍然是使用截值来选择使净收益达到最大的点。

Negative slope to reference curve



Oversampling and Asymmetric Costs
过度抽样和不对称成本

Rare Cases 罕见案例

Asymmetric costs/benefits typically go hand in hand with presence of rare but important class
不对称成本或者收益通常与罕见但重要的案例共存。

- Responder to mailing 对信件的应答
- Someone who commits fraud 某人从事诈骗
- Debt defaulter 无力偿债的借债人

- Often we oversample rare cases to give model more information to work with 我们经常使用过度抽样以便给模型分析更多的信息。

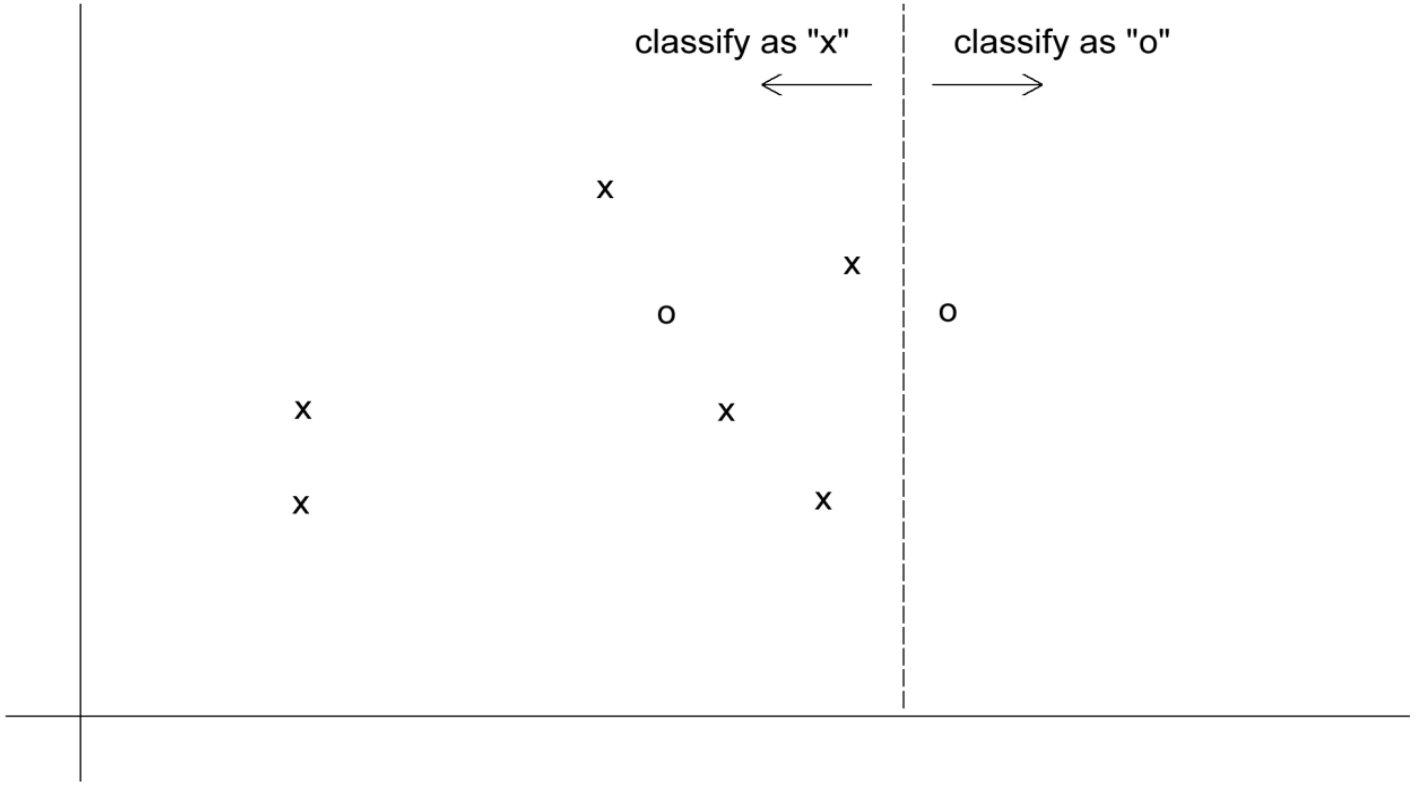
- Typically use 50% “1” and 50% “0” for training 通常用50%的结果为1的案例和50%结果为0的案例作为训练数据集。

Example 例子

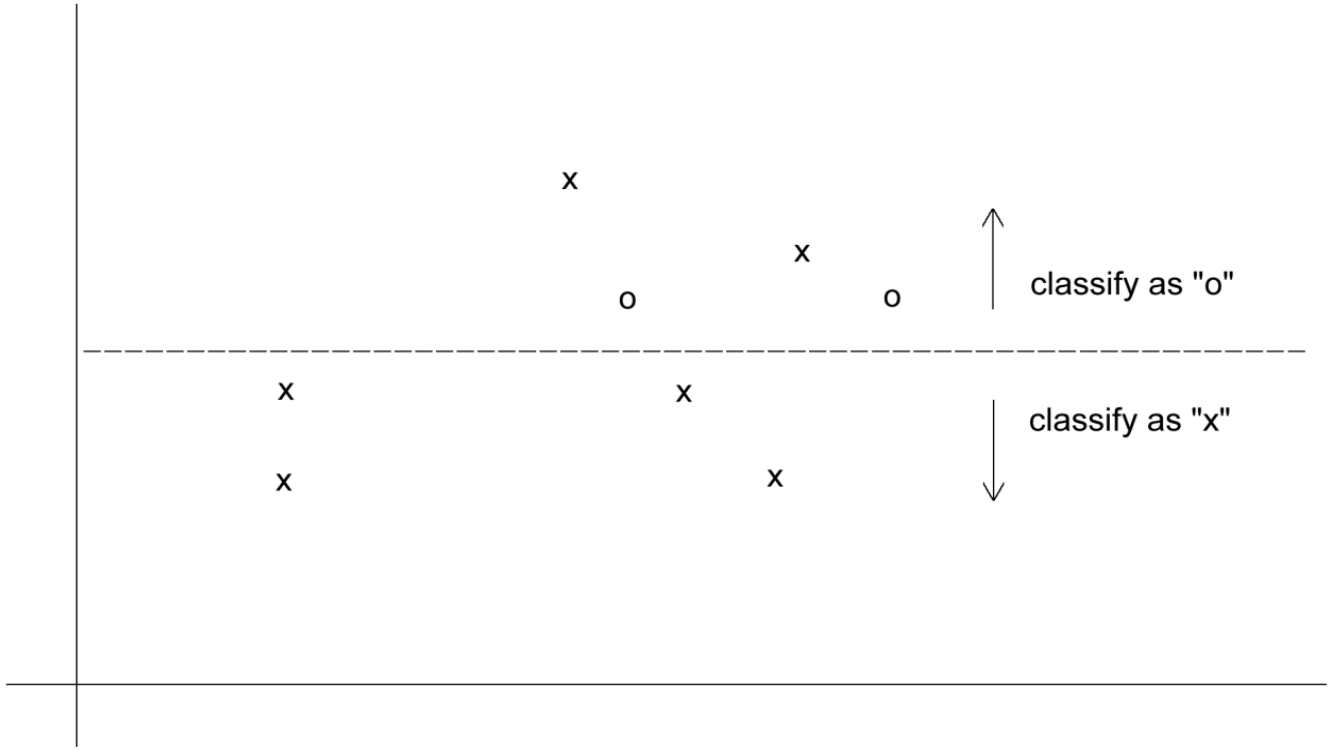
Following graphs show optimal classification under three scenarios: 下面的几张图显示三种情况下的最佳分类

- assuming equal costs of misclassification 假设误分类成本是一样的。
- assuming that misclassifying “o” is five times the cost of misclassifying “x” 假设将“o”误分类的成本是将“x”误分类的成本的5倍。
- Oversampling scheme allowing DM methods to incorporate asymmetric costs 使用过度抽样允许数据挖掘算法把不对称成本因素考虑进来。

Classification: equal costs 分类：相同误分类成本

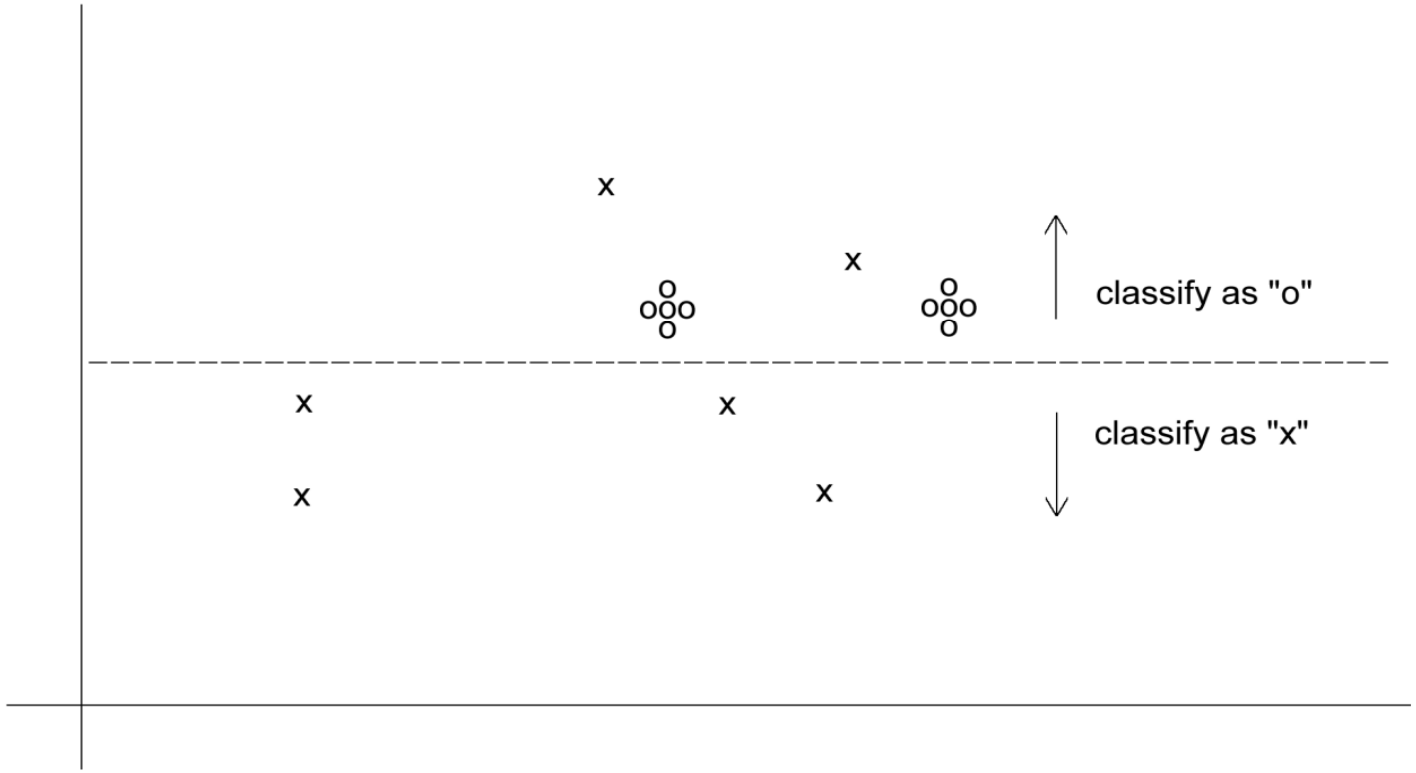


Classification: Unequal costs 分类：不同误分类成本



Oversampling Scheme 过度抽样方案

Oversample “o” to appropriately weight misclassification costs 过度对“o”进行抽样以便给将o误分类的成本更合适的权重。



An Oversampling Procedure 过度抽样的过程

1. Separate the responders (rare) from non-responders 将回应（罕见）的案子与不回应的案子分开。
2. Randomly assign half the responders to the training sample, plus equal number of non-responders 随机将一半做了回应的案子放入训练数据集，并随机从数据集中抽出相同数目的未做回应的案子。
3. Remaining responders go to validation sample 其余做了回应的案子放入验证数据集。
4. Add non-responders to validation data, to maintain original ratio of responders to non-responders 将未作回应的的数据放入验证数据集，保持该数据集中做了回应的案子比率与原数据集相同。
5. Randomly take test set (if needed) from validation 如果需要随机挑出检验数据集。

Classification Using Triage 用分流进行分类

Take into account a gray area in making classification decisions 做分类决策时考虑灰色区域

- Instead of classifying as C_1 or C_0, we classify as 不一定分成2类，还可以考虑第3类。
  - C_1
  - C_0
  - Can’t say
- The third category might receive special human review 第3类可能会被特别进行人工审查。

Evaluating Predictive Performance 评估预测模型的表现

Measuring Predictive error 度量预测错误

- Not the same as “goodness-of-fit” 与“拟合优度”不同。

- We want to know how well the model predicts new data, not how well it fits the data it was trained with 我们想知道模型预测新数据时有多好，而不是拟合训练数据集有多好。

- Key component of most measures is difference between actual y and predicted y (“error”) 度量的关键要素是预测值与实际值之间的偏差。

Some measures of error 对预测错误的一些度量指标

MAE or MAD: Mean absolute error (deviation) 误差绝对值的平均值

Gives an idea of the magnitude of errors

Average error 误差平均值

Gives an idea of systematic over- or under-prediction

MAPE: Mean absolute percentage error 误差百分比绝对值的平均值

RMSE (root-mean-squared-error): Square the errors, find their average, take the square root 对误差求平方，找到它们的平均值然后开方。

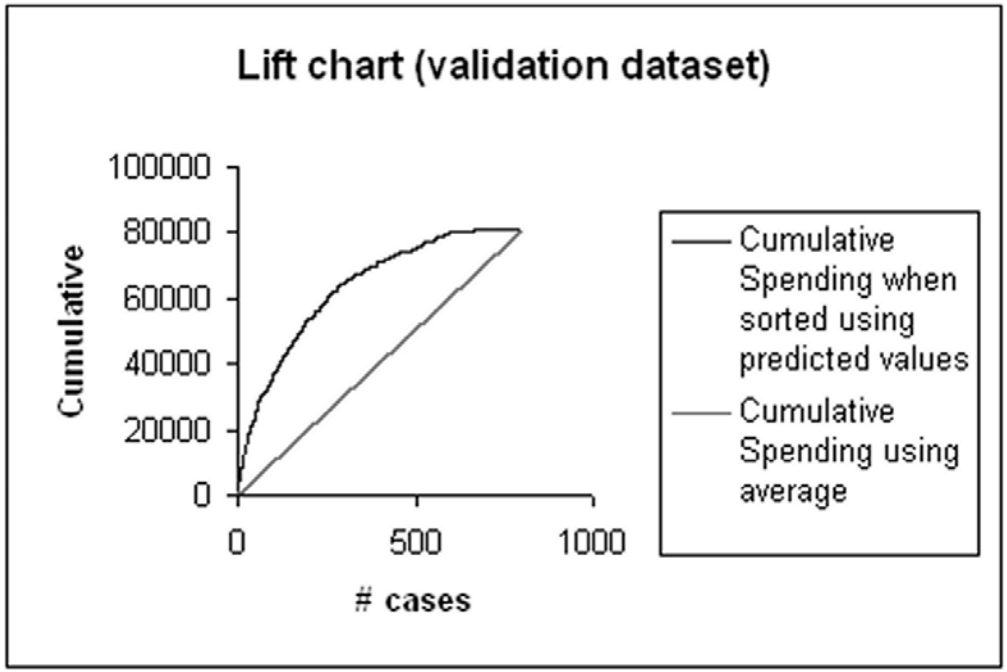
Total SSE: Total sum of squared errors 误差平方和

Lift Chart for Predictive Error 预测错误的提升图

Similar to lift chart for classification, except… 与分类的提升图类似，不同之处是

Y axis is cumulative value of numeric target variable (e.g., revenue), instead of cumulative count of “responses” Y轴是累积目标值（如销售收入），而不是累积“回应”的计数。

Lift chart example – spending



Summary 总结

- Evaluation metrics are important for comparing across DM models, for choosing the right configuration of a specific DM model, and for comparing to the baseline 评价矩阵的重要性体现在：比较不同的数据挖掘模型，选择某个数据挖掘模型的正确配置，以及与基准进行比较。
- Major metrics: confusion matrix, error rate, predictive error 主要的矩阵：混淆矩阵、错误率、预测错误。
- Other metrics when 使用其他矩阵当
  - one class is more important 一个类别更加重要时
  - asymmetric costs 误分类成本不对称
- When important class is rare, use oversampling 当重要的类别较罕见，使用过度抽样。
- In all cases, metrics computed from validation data 对所有的案例，矩阵是在验证数据集的基础上进行计算的。