

# Lecture2

## Judging the Performance of Classifiers

Int hisnot ew ew ill examinet he questiono fhow t o judget heus efulnesso fa c lassifiera nd howt o comparedi fferentc lassifiers. N otion lydow eha vea w idec hoice ofd iffereantt ypesof c lassifiers toc hoose frombut w ithine acht ypeo fc lassifierw eh ave manyopt ions. s ucha sho wmanyne arest neighborst ou se ina k -nearestn eighborsc lassifier, the minimumnu mberof c ases wes hould requirei n al eaf node ina treec lassifier, w hichs ubsets ofpr edictors tous ei na logistic regression model, a ndhow m anyhi ddenl ayer neurons tou sei na ne uraln et.

### AT wo-classC lassifier

Letus firstl ooka ta single classifier for twoc lassesw ithop tionss eta tc ertainv alues. The twoclass situation isc ertainlyt hem ostc ommona ndoc cursv eryf requentlyi npr actice. W ew ill extendou ra nalysis tom oret han twoc lasses later.

Ana tural criterionf or judgingt hepe rformanceof a c lassifier is thep robabilityt hati tm akesa misclassification. A c lassifiert hatm akesnoe rrorsw ouldbe p erfectbu tw edo not expectt obe able toc onstruct such classifiers int he realw orlddu et o“ noise” and tona tha vinga llt he informationn eededt opr eciselyc lassifyc ases. I st here am inimump r obabilityof misclassificationw es houldr equireof ac lassifier?

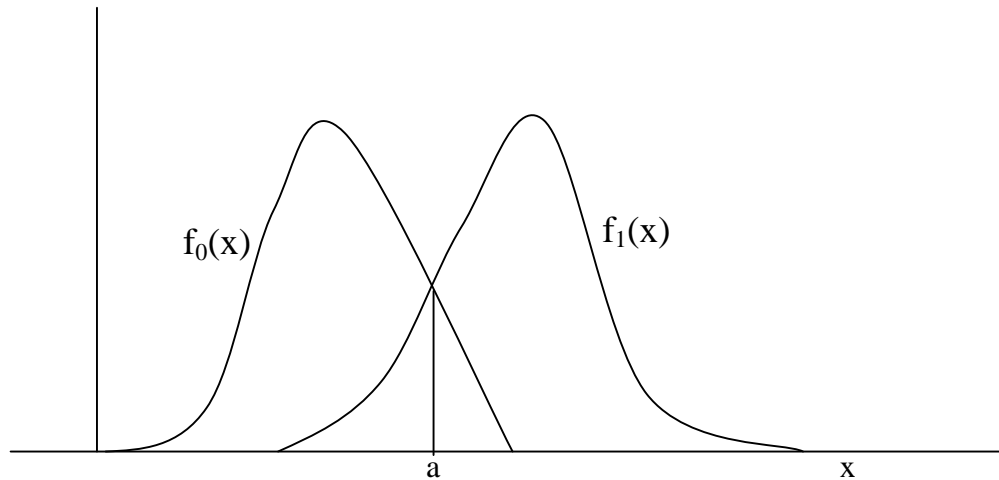
Suppose that the twoc lassesa rede noted by  $C_0$  a nd  $C_1$ . Let  $p(C_0)$  a nd  $p(C_1)$  be the apriori probabilityest hat ac asebe elongst o  $C_0$  a nd  $C_1$  r espectively. T hea priori probabilityi st he probabilityt hat ac asebe elongst oa c lassw ithouta nym orek nowledgea bout it thant hat itb elongstoa p opulationw here the proportiono f  $C_0$ 's is  $p(C_0)$  andt he proportionof  $C_1$ 's is  $p(C_1)$ . I nt his situationw ew illm inimizet hec hanceof am isclassificatione rrorby a ssigningc lass  $C_1$  t o the case if  $p(C_1) > p(C_0)$  and t o  $C_0$  ot herwise. T hep r obabilityof m aking am isclassificationw ould bet hem inimumof  $p(C_0)$  a nd  $p(C_1)$ . I fw ea reus ingm isclassification ratea sour c riterion any classifier thatus esp redictorv ariablesm ustha vea ne rrorr ate bettert han this.

Whati s thebe stpe rformancew ec ane xpectf roma c lassifier? C learlyt hem oret rainingda ta available toa classifiert he morea ccuratei tw illb e. S upposew e hada hug ea mountof training data, w ouldw et henbe ablet obu ild ac lassifier thاتم akesnoe rrors? The answeri sno. The accuracyof a c lassifierde pendsc riticallyonhow s eparatedt hec lasses arew ithr espectt o the predictorv ariables that the classifieru ses. We canu se thew ell-knownB ayes'f ormulaf rom probabilityt heoryt o derivet heb estpe rformancew ec ane xpectf roma c lassifierf ora g ivenst et ofpr edictorv ariables ifw e hada v eryl argea mountof trainingda ta. B ayes'f ormulaus est he distributionsof the decisionv ariables in the twoc lasseset og iveus a c lassifier that will havet he minimume rroramongsta llc lassifierst hatu se the samepr edictorv ariables. Thisc lassifier follows the MinimumE rrorB ayesR ule.

### BayesR ulef or MinimumE rror

Letus takea simples ituationw herew eha ve just one continuouspr edictorv ariablef or classification, s ay  $X$ .  $X$  isa randomv ariable, s incei t'sv aluede pendso nt hei ndividual casew e samplef romt hepopu lationc onsistingof a llpos siblec ases of thec lass tow hicht hec asebe longs. Suppose thatw eh avea v eryl arget rainingda ta set. Thent her elativef requencyhi stogramof t he variable  $X_i$  ne achc lassw ouldbe almost identicalt o thep r obabilityde nsityf unction( p.d.f.) of  $X$  for thatc lass. L etus assumet hatw eh avea hug ea mountof trainingda taa nds ow ek nowt he p.d.f.s accurately. T hesep. d.f.s are denoted  $f_0(x)$  a nd  $f_1(x)$  f or classes  $C_0$  a nd  $C_1$  i n  $F$  ig. 1 be low.

Figure 1



Now suppose we wish to classify a new object for which the value of  $X$  is  $x_0$ . Let us use Bayes' formula to predict the probability that the object belongs to class 1 conditional on the fact that it has a value of  $x_0$ . Applying Bayes' formula, the probability, denoted by  $p(C_1|X=x_0)$ , is given by:

$$p(C_1|X = x_0) = \frac{p(X = x_0|C_1)p(C_1)}{p(X = x_0|C_0)p(C_0) + p(X = x_0|C_1)p(C_1)}$$

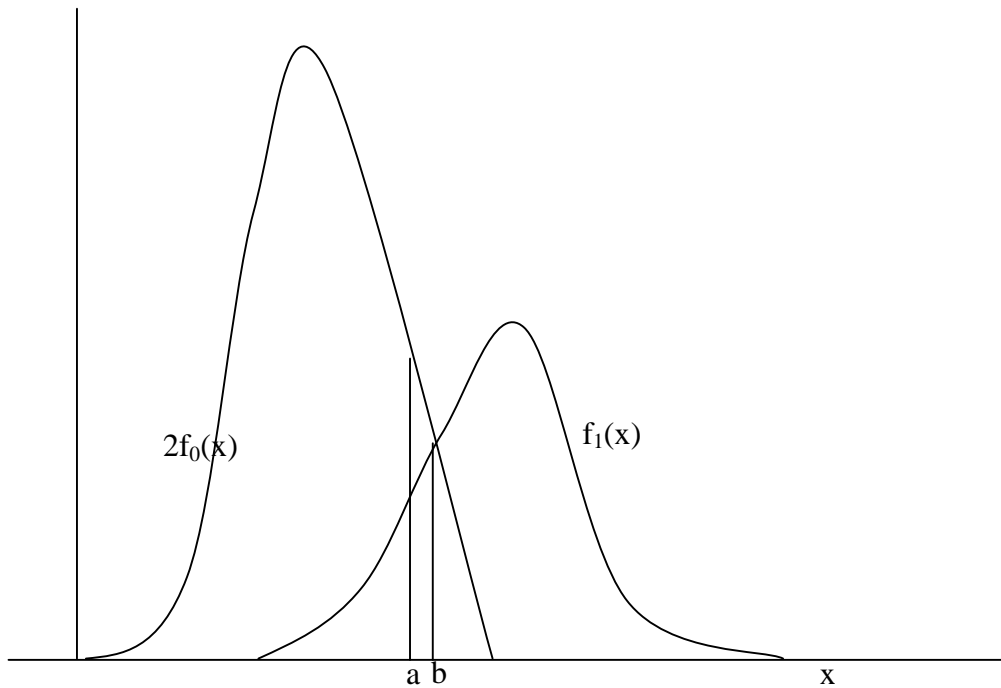
Writing this in terms of the density functions, we get

$$p(C_1|X = x_0) = \frac{f_1(x_0)p(C_1)}{f_0(x_0)p(C_0) + f_1(x_0)p(C_1)}$$

Notice that to calculate  $p(C_1|X=x_0)$  we need to know the prior probabilities  $p(C_0)$  and  $p(C_1)$ . Since there are only two possible classes, if we know  $p(C_1)$  we can always compute  $p(C_0)$  because  $p(C_0) = 1 - p(C_1)$ . The prior probability  $p(C_1)$  is the probability that an object belongs to  $C_1$  without any knowledge of the value of  $X$  associated with it. Bayes' formula enables us to update this a priori probability to the posterior probability, the probability of the object belonging to  $C_1$  after knowing that its  $X$  value is  $x_0$ .

When  $p(C_1) = p(C_0) = 0.5$ , the formulas show that  $p(C_1|X = x_0) > p(C_0|X = x_0)$  if  $f_1(x_0) > f_0(x_0)$ . This means that if  $x_0$  is greater than  $a$ , and we classify the object by belonging to  $C_1$  we will make a smaller misclassification error than if we were to classify it as belonging to  $C_0$ . Similarly, if  $x_0$  is less than  $a$ , and we classify the object by belonging to  $C_0$  we will make a smaller misclassification error than if we were to classify it as belonging to  $C_1$ . If  $x_0$  is exactly equal to  $a$ , we have a 50% chance of making an error in either classification.

Figure 2



What if the prior probabilities were not the same? Suppose  $C_0$  is twice as likely a priori as  $C_1$ . Then the formulas say that  $p(C_1|X = x_0) > p(C_0|X = x_0)$  if  $f_1(x_0) > 2 \times f_0(x_0)$ . Then the boundary value,  $b$ , for classification will be to the right of  $a$  as shown in Figure 2. This is intuitively what we would expect. If a class is more likely we would expect the cut-off to move in the direction that would increase the range over which it is preferred.

In general we will minimize the misclassification error if we classify  $C_1$  if  $p(C_1) \times f_1(x_0) > p(C_0) \times f_0(x_0)$ , and  $C_0$  otherwise. This rule holds even when  $X$  is a vector consisting of several components, each of which is a random variable. In the remainder of this section we will assume that  $X$  is a vector.

An important advantage of Bayes' Rule is that, as a by-product of classifying a case, we can compute the conditional probability that the case belongs to each class. This is a significant advantage.

First, we can set the probability as a "score" for each case that we are classifying. This score enables us to rank cases that we have predicted to belong to a class in order of confidence that we have made a correct classification. This capability is important in developing a lift curve (explained later) that is important for many practical data mining applications.

Second, it enables us to compute the expected profit or loss for a given case. This gives us a better decision criterion than the misclassification error when the losses are different for the two classes.

Practical assessment of a classifier using misclassification error as the criterion

In practice, we can estimate  $p(C_1)$  and  $p(C_0)$  from the data used to build the classifier by simply computing the proportion of cases that belong to each class. Of course, these are estimates and they can be incorrect, but they are often good enough data to estimate the true probabilities. Sometimes, when we have a large public data set such as census data, the estimates will be reliable. However, in most practical business settings we will not know  $f_1(x)$  and  $f_0(x)$ . If we want to apply Bayes' Rule we will need to estimate these density functions in some way. Many classification methods can be interpreted as estimating the density functions<sup>1</sup>. In practice, it is almost always better to use a vector machine like the task difficult and subject to the curse of dimensionality rather than discussing the k-Nearest Neighbors technique.

To obtain an estimate of classification error, let us suppose that we have partitioned our data set into training and validation data sets by random selection of cases. Let us assume that we have constructed a classifier using the training data. When we apply it to the validation data, we will classify each case into  $C_0$  or  $C_1$ . The resulting misclassification error can be displayed in a confusion table, with rows and columns corresponding to the true and predicted classes respectively. We can summarize our results in a confusion table in a similar fashion. The resulting confusion table will not give us a point estimate of the misclassification rate due to overfitting. However, it will be useful to signal overfitting when the substantially lower misclassification rates than the confusion table for validation data.

Confusion Table (Validation Cases)	Predicted Class	
	$C_0$	$C_1$
$C_0$	True Negatives (Number of correctly classified cases that belong to $C_0$ )	False Positives (Number of cases incorrectly classified as $C_1$ that belong to $C_0$ )
$C_1$	False Negatives (Number of cases incorrectly classified as $C_0$ that belong to $C_1$ )	True Positives (Number of correctly classified cases that belong to $C_1$ )

If we denote the number in the cell in row  $i$  and column  $j$  by  $N_{ij}$ , the estimated misclassification rate is  $E_{rr} = (N_{01} + N_{10}) / N_{val}$  where  $N_{val} \equiv (N_{00} + N_{01} + N_{10} + N_{11})$ , or the total number of cases in the validation data set. If  $N_{val}$  is reasonably large, our estimate of the misclassification

<sup>1</sup> There are classifiers that focus on simply finding the boundary between the regions to predict each class without being concerned with estimating the density of cases within each region. For example, Support Vector Machine classifiers have this characteristic.

ratei spr obablyqui tea ccurate. Wec anc omputea c onfidencei ntervalf orE rru singt hes tandard formulaf ore stimatinga po pulationpr oportionf roma r andoms ample.

The table below gives a nice idea of how the accuracy of the estimate varies with  $N_{val}$ . The column headings are values of the misclassification rate and the rows give the corresponding accuracy in estimating the misclassification rate as measured by the half-width of the confidence interval at the 99% confidence level. For example, if we think that the true misclassification rate is likely to be around 0.05 and we want to be 99% confident that the error is within  $\pm 0.01$  of the true misclassification rate, we need to have a validation data set with 3,152 cases.

	0.01	0.05	0.10	0.15	0.20	0.30	0.40	0.50
$\pm 0.025$	250	504	956	1,354	1,699	2,230	2,548	2,654
$\pm 0.010$	657	3,152	5,972	8,461	10,617	13,935	15,926	16,589
$\pm 0.005$	2,628	12,608	23,889	33,842	42,469	55,741	63,703	66,358

Notet hatw ea rea ssumingt hat thec ost( orbe nefit)of makingc orrectc lassificationsi sz ero. A t firstg lance, thism ays eem incomplete. A fter a ll, theb enefit (negativec ost)o fc orrectly classifyinga buy era s abuy erw oulds eems ubstantial. And, i no ther c ircumstances( e.g. calculatingt hee xpected profitf romha vinga ne wm ailingl ist ), itw illbe appropriatet oc onsider thea ctualn et dollari mpact ofc lassifyinge achc aseon the list. H ere, h owever, w ea re attempting toa ssesst hev alue ofa c lassifieri n terms of m isclassifications, so itg reatly implifiesm atters if wec anc apture allc ost/benefiti nformation in them isclassificationc ells. S o, i nstead of recording theb enefitof correctl y c lassifyinga buy er, w e record thec ost o ff ailingt o classifyhi ma sa buyer. I ta mountst ot hes amet hing. I n fact the costs w ea re usinga ret heap oportunit yc osts.

### Asymmetric misclassification costs and Bayes' Risk

Up to this point we have been using the error rate as the criterion for judging the efficacy of a classifier. However, there are circumstances when this measure is not appropriate. Sometimes the cost of misclassifying a case is long-term or more serious than for the other class. For example, misclassifying a householder as likely to respond to a sales offer when it belongs to the class that would respond is a greater opportunity cost than the converse error. In such a scenario, minimizing misclassification error as a criterion can be misleading. Consider the situation where the sales offer is accepted by 1% of the household sample. If a classifier simply classifies every household as a non-responder it will have an error rate of only 1% but will be useless in practice. A classifier that misclassifies 30% of buying households as non-buyers and 2% of non-buyers as buyers would have a higher error rate but would be better if the profit from a sale is substantially higher than the cost of sending out an offer. In these situations, if we have estimates of the cost of both types of misclassification, we can use the confusion table to compute the expected cost of misclassification for each class in the validation data. This enables us to compare different classifiers using opportunity cost as the criterion. This may suffice for some situations, but a better method would be to construct classification rules (and hence the misclassification rates) to reflect the asymmetric costs. In fact, there is a Bayes' Risk classifier for this situation which gives rules that are optimal for minimizing the expected opportunity loss from misclassification. This classifier is known as the Bayes' Risk classifier and the corresponding minimum expected opportunity cost of misclassification is known as the Bayes' Risk. The Bayes' Risk classifier employs the following classification rule:

Classify a case as belonging to  $C_1$  if  $p(C_1) \times f_1(x_0) \times C(0|1) > p(C_0) \times f_0(x_0) \times C(1|0)$ , and to  $C_0$  otherwise. Here  $C(0|1)$  is the opportunity cost of misclassifying a  $C_1$  case as belonging to  $C_0$  and  $C(1|0)$  is the opportunity cost of misclassifying a  $C_0$  case as belonging to  $C_1$ . Note that the opportunity cost of correct classification for either class is zero. Notice also that this rule reduces to the Minimum Error Bayes' Rule when  $C(0|1) = C(1|0)$ .

Again, as we rarely know  $f_1(x)$  and  $f_0(x)$ , we cannot construct this classifier in practice. Nonetheless, it provides us with a method for dealing with various classifiers we construct from minimizing expected opportunity cost to estimate them. The method is to use stratified sampling instead of random samples to construct the training set to reflect the relative cost of making the two types of misclassification errors.

### Stratified sampling to detect bias in classification accuracy

The basic idea in using stratified sampling is to oversample the minority class to increase the weight given to the minority class in the loss function. It is well known that the opportunity cost of misclassifying a class  $C_1$  is greater than the cost of misclassifying a class  $C_0$  if the cost of misclassifying a class  $C_1$  is greater than the cost of misclassifying a class  $C_0$ . By virtue of this oversampling, the training data will automatically give the classifier more accurate information about the minority class than the majority class. Most of the time the class that has the higher misclassification cost will be the less frequent class (for example, fraudulent cases). In this situation rather than reduce the number of fraudulent transactions, a good thumb rule of ten used in practice is to sample a number of cases from each class. This is a very commonly used option since it ends up producing rules that require fewer features relative to the overall number of features.

### Generalization to more than two classes

All the comments made above about two-class classification extend readily to classification into more than two classes. Let us suppose we have  $k$  classes  $C_0, C_1, C_2, \dots, C_{k-1}$ . Then Bayes' formula gives us:

$$p(C_j | X = x_0) = \frac{f_j(x_0)p(C_j)}{\sum_{i=0}^{k-1} f_i(x_0)p(C_i)}$$

The Bayes' rule for minimum error cost classification is to classify a case as belonging to  $C_j$  if  $p(C_j) \times f_j(x_0) \geq \max_{i=0,1,\dots,k-1} p(C_i) \times f_i(x_0)$ .

The confusion table has  $k$  rows and  $k$  columns. The opportunity cost associated with the diagonal cells is always zero. If the cost matrix is symmetric the Bayes' rule classifier follows the rule: Classify a case as belonging to  $C_j$

$$\text{if } p(C_j) \times f_j(x_0) \times C(\sim j | j) \geq \max_{i \neq j} p(C_i) \times f_i(x_0) \times C(\sim i | i).$$

where  $C(\sim j | j)$  is the cost of misclassifying a case that belongs to  $C_j$  as belonging to class  $C_i, i \neq j$ .

Lift Charts for two-class classifiers

Often in practice, opportunity costs are not taken into account in the decision-making process. In such cases, when the classifier gives a probability of belonging to each class and not just a binary (or "hard") classification of  $C_1$  or  $C_0$ , we can save very useful information by using the lift chart. The lift chart is a popular technique in direct marketing. The input required to construct a lift chart is a validation dataset that has been "scored" by applying the probability predicted by a classifier to each case. In fact, we can use a classifier that does not predict probabilities but gives scores that are used to rank cases in order of how likely they are to belong to one of the classes.

Example: Boston Housing (Two classes)

Let us fit a logistic regression model to the Boston Housing data. We fit a logistic regression model to the training data (304 randomly selected cases) with all the 13 variables available in the data set as predictor variables and with the variable HICLASS (highly polluted neighborhood) as the dependent variable. The model coefficients are applied to the validation data (the remaining 202 cases in the data set). The first three columns of XLMiner output or the first 30 cases in the validation data are shown below.

	Predicted Log-odds of Success	Predicted Prob. of Success	Actual Value of HICLASS
1	3.5993	0.9734	1
2	-6.5073	0.0015	0
3	0.4061	0.6002	0
4	-14.2910	0.0000	0
5	4.5273	0.9893	1
6	-1.2916	0.2156	0
7	-37.6119	0.0000	0
8	-1.1157	0.2468	0
9	-4.3290	0.0130	0
10	-24.5364	0.0000	0
11	-21.6854	0.0000	0
12	-19.8654	0.0000	0
13	-13.1040	0.0000	0
14	4.4472	0.9884	1
15	3.5294	0.9715	1
16	3.6381	0.9744	1
17	-2.6806	0.0641	0
18	-0.0402	0.4900	0
19	-10.0750	0.0000	0
20	-10.2859	0.0000	0
21	-14.6084	0.0000	0
22	8.9016	0.9999	1
23	0.0874	0.5218	0
24	-6.0590	0.0023	1
25	-1.9183	0.1281	1
26	-13.2349	0.0000	0

27	-9.6509	0.0001	0
28	-13.4562	0.0000	0
29	-13.9340	0.0000	0
30	1.7257	0.8489	1

These cases are shown below in descending order of the predicted probability of being a HIGH CLASS case.

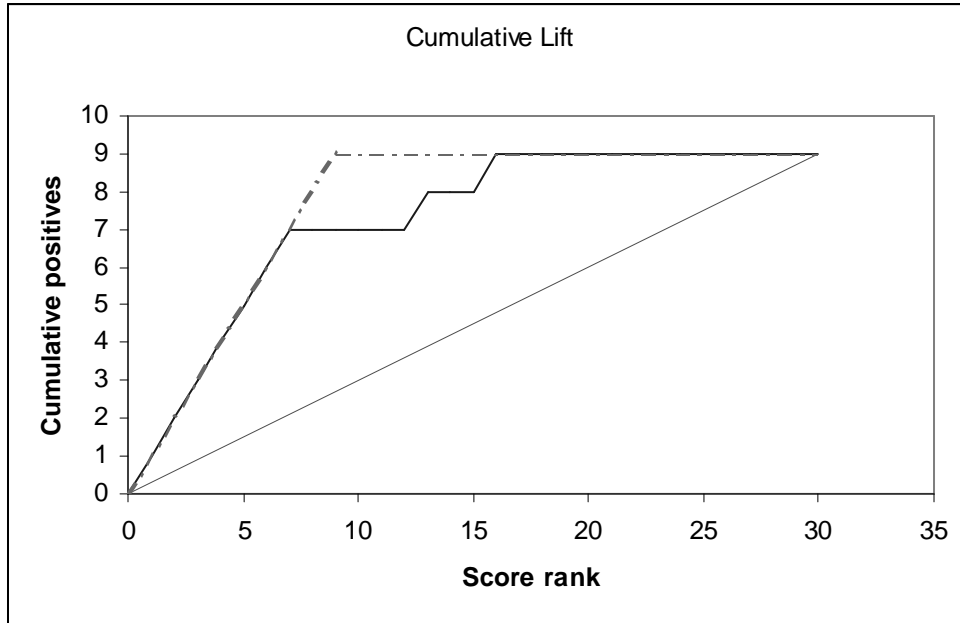
	Predicted Log-odds of Success	Predicted Prob. of Success	Actual Value of HICLASS
22	8.9016	0.9999	1
5	4.5273	0.9893	1
14	4.4472	0.9884	1
16	3.6381	0.9744	1
1	3.5993	0.9734	1
15	3.5294	0.9715	1
30	1.7257	0.8489	1
3	0.4061	0.6002	0
23	0.0874	0.5218	0
18	-0.0402	0.4900	0
8	-1.1157	0.2468	0
6	-1.2916	0.2156	0
25	-1.9183	0.1281	1
17	-2.6806	0.0641	0
9	-4.3290	0.0130	0
24	-6.0590	0.0023	1
2	-6.5073	0.0015	0
27	-9.6509	0.0001	0
19	-10.0750	0.0000	0
20	-10.2859	0.0000	0
13	-13.1040	0.0000	0
26	-13.2349	0.0000	0
28	-13.4562	0.0000	0
29	-13.9340	0.0000	0
4	-14.2910	0.0000	0
21	-14.6084	0.0000	0
12	-19.8654	0.0000	0
11	-21.6854	0.0000	0
10	-24.5364	0.0000	0
7	-37.6119	0.0000	0

First, we need to set a cutoff probability value, above which we will consider a case to be a positive ("1"), and below which we will consider a case to be a negative ("0"). For any given cutoff level, we can use the sorted table to compute a confusion table for a given cutoff probability. For example, if we use a cutoff probability level of 0.400, we will predict 10 positives (7 true positives and 3 false positives); we will also predict 20 negatives (18 true negatives and 2 false negatives). For each cutoff level, we can calculate the appropriate confusion table. Instead of looking at a large number of confusion tables, it is much more convenient to look at the cumulative lift curve (sometimes called a gain chart) which

summarizes all the information in these multiple confusion tables into a graph. The graphs are constructed with the cumulative number of successes (in descending order of probability) on the x-axis and the cumulative number of true positives on the y-axis as shown below.

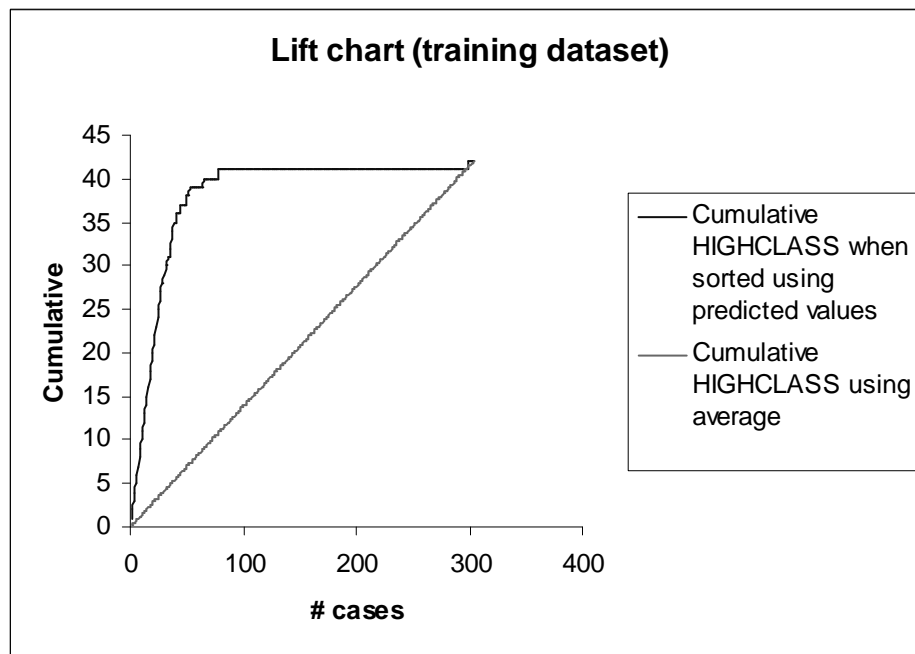
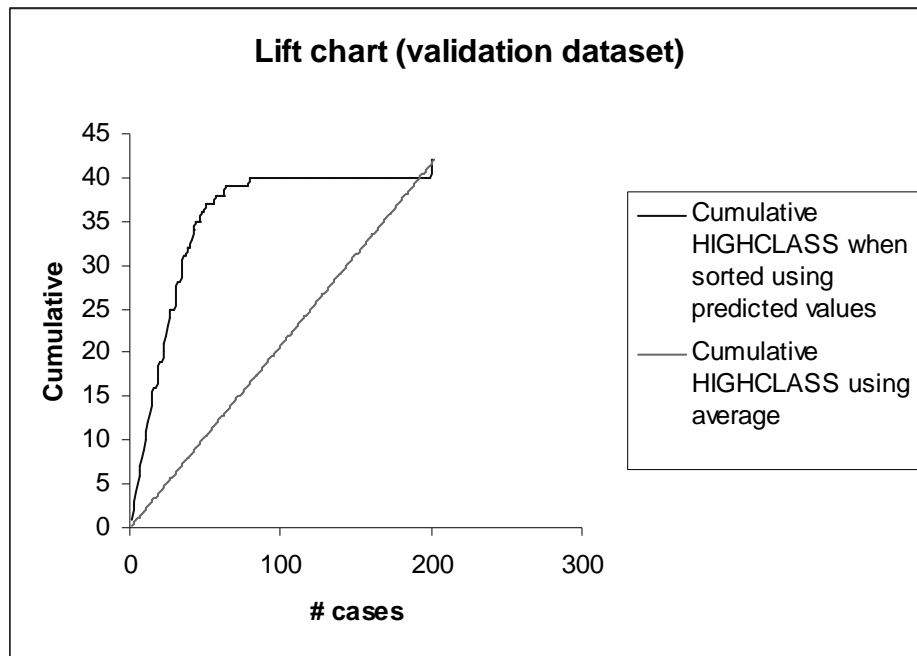
Probability Rank	Predicted Prob. of Success	Actual Value of HICLASS	cumulative Actual Value
1	0.9999	1	1
2	0.9893	1	2
3	0.9884	1	3
4	0.9744	1	4
5	0.9734	1	5
6	0.9715	1	6
7	0.8489	1	7
8	0.6002	0	7
9	0.5218	0	7
10	0.4900	0	7
11	0.2468	0	7
12	0.2156	0	7
13	0.1281	1	8
14	0.0641	0	8
15	0.0130	0	8
16	0.0023	1	9
17	0.0015	0	9
18	0.0001	0	9
19	0.0000	0	9
20	0.0000	0	9
21	0.0000	0	9
22	0.0000	0	9
23	0.0000	0	9
24	0.0000	0	9
25	0.0000	0	9
26	0.0000	0	9
27	0.0000	0	9
28	0.0000	0	9
29	0.0000	0	9
30	0.0000	0	9

The cumulative lift chart is shown below.



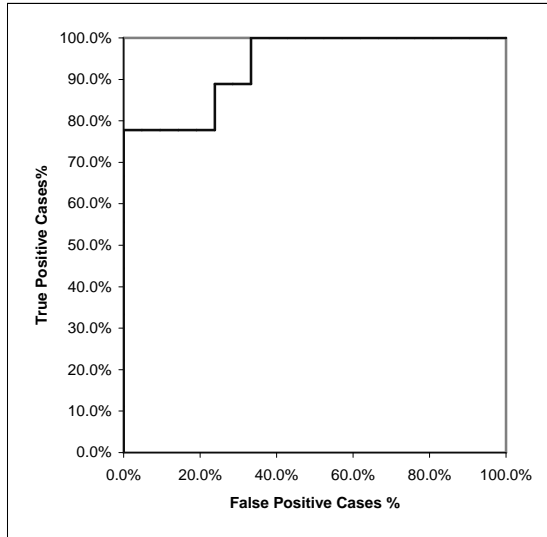
The line joining the points (0,0) to (30,9) is a reference line. It represents the expected number of positives we would predict if we did not know anything about the cases. It provides a benchmark against which we can see performance of the model. If we had to choose 10 neighborhoods as high risk, the lift curve tells us that we could be right about 7 of them. If we simply select 10 cases at random we expect to be right for  $10 \times 9 / 30 = 3$  cases. The model gives us a "lift" in predicting high risk of  $7 / 3 = 2.33$ . The lift will vary with the number of cases we choose to act on. A good classifier will give us a high lift when we act on only a few cases (i.e. use the prediction for the most serious cases). As we include more cases the lift will decrease. The lift curve for the best possible classifier is shown as a broken line.

XLMiner automatically creates lift charts from probabilities predicted by logistic regression for both training and validation data. The charts created for the full Boston Housing data are shown below.



It is worth mentioning that a curve that captures the same information as the lift curve is a slightly different manner is also popular in data mining applications. This is the ROC (short for Receiver Operating Characteristic) curve. It is just the same variable on the x-axis as the lift curve (but expressed as a percentage of the maximum) and on the y-axis it shows the false positives (also expressed as a percentage of the maximum) for different cut-off levels.

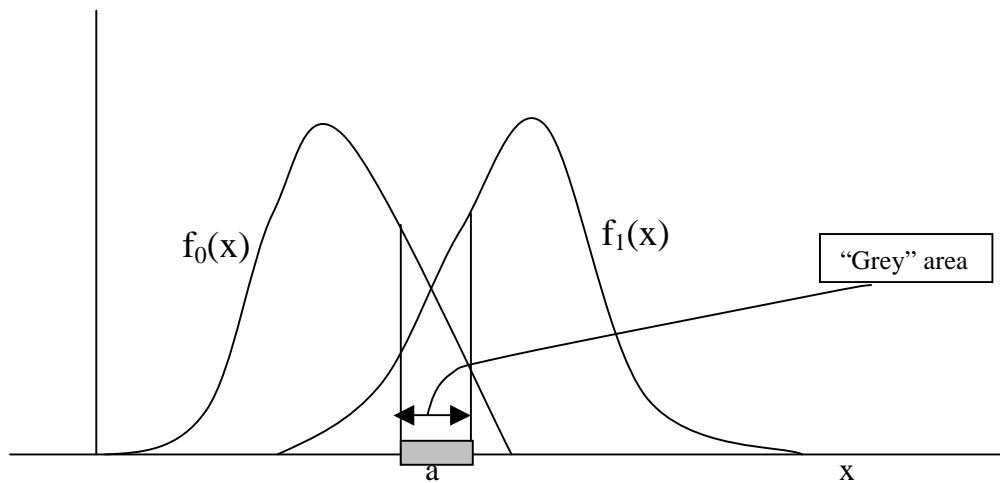
The ROC curve for the 30 case sample above is shown below.



### Classification using a Triages strategy

In some cases it is useful to have a “can’t say” option for the classifier. In a two-class situation this means that for each case one of three predictions. The case belongs to  $C_0$ , or the case belongs to  $C_1$ , or we cannot make a prediction because there isn’t enough information to confidently pick  $C_0$  or  $C_1$ . Cases that the classifier cannot classify are subjected to closer scrutiny either by using expert judgment or by enriching the set of predictor variables by gathering additional information that is perhaps more difficult or expensive to obtain. This is analogous to the strategy of triage that is often employed in a hospital. The wounded are classified into those who are able to walk or retreat, those who are too ill to retreat and eventually treated under the prevailing conditions, and those who are likely to become well enough or retreat if given medical attention. A nice example is in processing credit card transactions where a classifier may be used to identify clearly legitimate cases and the obviously fraudulent ones while referring the remaining cases to a human decision-maker who may look up a data set of former judgment. Since the vast majority of transactions are legitimate, such a classifier would substantially reduce the burden on human experts.

To gain some insight into the workings of such a strategy let us revisit the simple two-class, one predictor variable, classifier that we examined at the beginning of this chapter.



Clearly the grey area represents the uncertainty in classification. A threshold  $t$  of the conditional probabilities of belonging to the classes is one. A sensible rule to define the grey area is the set of  $x$  values such that:

$$t > \frac{p(C_1) \times f_1(x_0)}{p(C_0) \times f_0(x_0)} > 1/t$$

where  $t$  is a threshold for the ratio. A typical value of  $t$  may not be in the range 1.05 to 1.2.