Classification

Evaluation: Accuracy, error, precision, recall, F_1

`http://www.kamperh.com/`

Classification accuracy and error

$$\text{Accuracy} = \frac{\sum_{n=1}^{N} \mathbb{I}\left\{y^{(n)} = \hat{y}^{(n)}\right\}}{N}$$

$$\text{Error} = 1 - \text{Accuracy}$$

Binary classification:

$$\hat{y} = \begin{cases} 1 & \text{if } f(\underline{x}; \underline{w}) \geq 0.5 \\ 0 & \text{if } f(\underline{x}; \underline{w}) < 0.5 \end{cases}$$

Multiclass classification:

$$\hat{y} = \arg\max_{k=1}^{K} f_k(\underline{x}; \underline{W})$$

- Often useful as single numbers to summarise and compare system performance.
- But can also, unfortunately, be "skewed" in some cases.
- For instance when one class occurs a lot more often than others.

Further motivation for more metrics

- Sometimes we might just be more interested in some classes than others.
- For instance, in binary classification we might have that $y = 1$ is a rare class that we are specifically interested in detecting.
- We might even be okay with accidentally classifying input that is $y = 0$ as positive, as long as all the true $y = 1$ cases are detected.
- In other cases, it might be more important to be absolutely sure that when we make a positive prediction, that the true label is actually $y = 1$, even if we then accidentally miss some $y = 1$ cases and classify them as negative.
- Accuracy and error measure the importance of all classes equally. We therefore need metrics that break down performance more carefully.

Confusion matrix

		Actual class: 0	Actual class: 1
Predicted class	0	True negative	False negative
	1	False positive	True positive

Precision:

Of items classified as $y = 1$, what fraction is actually $y = 1$?

E.g. of all patients predicted to have cancer, how many actually do?

$$\text{Precision} = \frac{\text{No. true positives}}{\text{No. predicted positives}} = \frac{TP}{FP + TP}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Recall: Also called "sensitivity"

Of items that are actually $y = 1$, what fraction did we correctly predict as $y = 1$?

E.g. of all patients having cancer, how many are classified as having cancer?

$$\text{Recall} = \frac{\text{No. true positives}}{\text{No. actual positives}} = \frac{TP}{FN + TP}$$

Predicted class \ Actual class	0	1
0	True negative	False negative
1	False positive	True positive

Recall

Precision

F_1-score:

Recall and precision are combined by taking the harmonic mean:

$$F_1 = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Example: Predicting when someone defaults

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

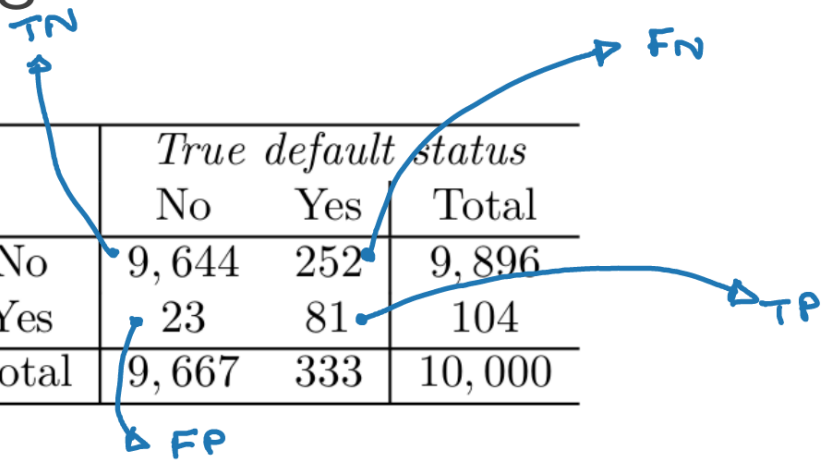

TABLE 4.4. *A confusion matrix compares the LDA predictions to the true default statuses for the* $10,000$ *training observations in the* `Default` *data set.*

Calculate accuracy, precision, recall and F_1 scores for:

1. The LDA classifier in the above table.
2. A classifier applied to the same data, but always predicting $\hat{y} = 0$.

Source: G. James, D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning*.

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

TN, FN, FP, TP

(i) LDA classifier:

$$\text{Accuracy} = \frac{TN + TP}{N} = \frac{9644 + 81}{10000} = 97.25\%.$$

$$\text{Prec.} = \frac{TP}{TP + FP} = \frac{81}{81 + 23} = 77.88\%.$$

$$\text{Rec.} = \frac{TP}{TP + FN} = \frac{81}{252 + 81} = 24.32\%.$$

(ii) A negative classifier

$f(\underline{x}; \underline{w}) = 0$

```
Code:
def f_predict(x):
    return 0
```

$$\text{Accuracy} = \frac{TN + TP}{N} = \frac{9667}{10000} = 96.67\%.$$

$$\text{Prec.} = \frac{TP}{TP + FP} = \frac{0}{0 + 0} = \text{NaN}$$

$$\text{Rec.} = \frac{TP}{TP + FN} = \frac{0}{0 + 333} = 0\%.$$

$\hat{y}$ \ y	0	1
0	TN 9667	FN 333
1	FP 0	TP 0

Source: G. James, D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning*.

Trading off precision and recall

Binary classification prediction:

$$\hat{y} = \begin{cases} 1 & \text{if } f(\mathbf{x}; \mathbf{w}) \geq 0.5 \\ 0 & \text{if } f(\mathbf{x}; \mathbf{w}) < 0.5 \end{cases}$$

Two examples:

1. High-precision miscue detection in a reading tutor.
2. Early, cheap TB scanner in a hospital (high-recall)

Binary classification prediction with threshold α:

$$\hat{y} = \begin{cases} 1 & \text{if } f(\mathbf{x}; \mathbf{w}) \geq \alpha \\ 0 & \text{if } f(\mathbf{x}; \mathbf{w}) < \alpha \end{cases}$$

Metrics for multiple classes

- Above we used precision, recall, F_1 to evaluate binary classification.
- It can also be extended to multiple classes. Let's look at one approach.
- Calculate precision and recall by treating each class in turn as the positive class.
- Then average the precisions and recalls (unweighed) across the classes.
- This gives the *macro precision* and *macro recall*.