Machine Learning,
Machine Learning (extended)

5 – Supervised Learning:
Evaluation Metrics for Classification

Kashif Rajpoot

k.m.rajpoot@cs.bham.ac.uk

School of Computer Science

University of Birmingham

Outline

- Performance evaluation
- 0/1 loss
- Classification accuracy
- Confusion matrix
  - True positive
  - True negative
  - False positive
  - False negative
- Sensitivity
- Specificity
- ROC analysis
  - Area under curve
- Hold-out validation
- Cross-validation
  - K-fold
  - Leave-one-out
- Repeated cross-validation

Performance evaluation

- How to assess a classification algorithm?
- How to choose?
  - Classification algorithm?
  - Algorithm parameters?

0/1 loss

- 0/1 loss: proportion of times classifier is wrong
- Consider a set of classifier label predictions $t_1, t_2, \ldots, t_N$ and *ground truth* target labels $t_1^*, t_2^*, \ldots, t_N^*$
- Mean 0/1 loss can be computed as:
  - $\frac{1}{N}\sum_{n=1}^{N} \delta(t_n \neq t_n^*)$
- For a particular test sample prediction t_n:
  - $\delta(t_n \neq t_n^*) = 1$
  - $\delta(t_n = t_n^*) = 0$
- The lower the 0/1 loss, the better
- Advantages
  - Simple
  - Suitable for binary or multi-class classification

0/1 loss

- Disadvantage
  - Suffers from class imbalance

- Imagine we're building a classifier to detect a rare disease
  - Consider only 1% of population is diseased
  - $t = 1$, for diseased
  - $t = 0$, for healthy
- What if algorithm always predicts $t = 0$?
- Accuracy will be 99%, but the classification algorithm is rubbish

Classification accuracy

- A slight variant of 0/1 loss, computed as percentage accuracy
- Consider a set of classifier label predictions $t_1, t_2, \ldots, t_N$ and *ground truth* target labels $t_1^*, t_2^*, \ldots, t_N^*$
- Mean classification accuracy can be computed as:
  - $100 * \frac{1}{N} \sum_{n=1}^{N} \delta(t_n = t_n^*)$
- For a particular test sample prediction t_n:
  - $\delta(t_n \neq t_n^*) = 0$
  - $\delta(t_n = t_n^*) = 1$
- The higher the classification accuracy, the better
- Disadvantage
  - Suffers from class imbalance, similar to 0/1 loss

Confusion matrix (CM)

- Shows the simple count of correctly (and wrongly) classified samples by the classifier
- Consider a binary classification problem with 20 diseased and 80 healthy test samples

		Actual class label		Total
		Diseased	Healthy	
Predicted class label	Diseased	15	4	19
	Healthy	5	76	81
	Total	20	80	

- Variety of metrics can be estimated from CM
- Suitable for binary and multi-class classification
  - Particularly useful for multi-class problems
  - Specifically indicates problems with an individual class

Confusion matrix

- Example: classify ~7000 test documents in 20 classes (newsgroups data)
  - Too similar classes?
  - Need more data?

Predicted class \ True class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	18	18	19	20
1	242	3	3	0	1	0	0	1	0	4	2	0	2	10	4	7	1	12	7	47
2	0	296	33	8	8	42	9	1	1	0	0	4	18	7	8	2	0	1	1	3
3	0	6	209	15	9	8	4	0	0	0	0	1	0	1	0	1	0	0	0	0
4	0	12	60	303	36	12	46	2	0	1	0	1	28	3	0	0	0	0	0	0
5	0	8	10	22	277	2	21	0	0	1	0	2	7	0	0	1	1	0	0	0
6	1	21	30	2	2	304	0	1	0	3	0	1	3	0	1	2	0	0	1	0
7	0	1	0	5	5	1	235	5	1	2	0	1	1	0	0	0	1	0	0	0
8	0	3	1	6	4	0	31	356	25	3	1	0	9	4	0	0	2	2	1	0
9	0	2	2	0	1	2	5	4	353	1	0	0	2	0	1	0	1	1	0	1
10	0	0	2	0	1	1	0	2	2	348	4	0	0	1	0	0	1	1	0	0
11	1	0	1	1	0	0	1	0	0	16	382	0	1	0	1	0	1	1	0	0
12	1	16	16	5	4	10	3	1	1	2	0	360	45	0	4	1	3	4	3	1
13	1	4	1	24	16	0	9	5	1	2	0	3	260	3	4	0	0	0	0	0
14	2	3	4	0	8	0	2	0	1	0	2	2	6	324	4	1	1	0	3	3
15	3	7	4	1	2	3	3	2	0	0	1	0	4	3	336	0	2	0	7	5
16	39	4	5	0	0	1	3	1	1	3	2	2	5	17	4	376	3	7	2	68
17	4	0	0	0	3	1	1	5	4	1	0	9	0	3	1	3	325	3	95	19
18	7	1	0	0	0	1	3	1	2	2	1	0	2	6	2	1	2	325	4	5
19	7	2	9	0	6	2	5	8	5	8	4	8	0	10	21	1	16	19	185	7
20	10	0	1	0	0	0	1	0	0	0	0	1	0	1	1	2	4	0	1	92

Confusion matrix (CM)

		Actual class label		Total
		Diseased	Healthy	
Predicted class label	Diseased	TP	FP	TP+FP
	Healthy	FN	TN	FN+TN
	Total	TP+FN	FP+TN	

- True positive (TP): diseased samples are classified as diseased
  - Number of test samples with $t_n^* = 1$ that are classified as $t_n = 1$
- True negative (TN): healthy samples are classified as healthy
  - Number of test samples with $t_n^* = 0$ that are classified as $t_n = 0$
- False positive (FP): healthy samples are classified as diseased
  - Number of test samples with $t_n^* = 0$ that are classified as $t_n = 1$
- False negative (FN): diseased samples are classified as healthy
  - Number of test samples with $t_n^* = 1$ that are classified as $t_n = 0$

Sensitivity and Specificity

- Sensitivity: proportion of diseased samples that are classified as diseased
  - The higher, the better

$$Sen = \frac{TP}{TP + FN} = \frac{TP}{All\ Positive}$$

- Specificity: the proportion of healthy samples that are classified as healthy
  - The higher, the better

$$Spec = \frac{TN}{TN + FP} = \frac{TN}{All\ Negative}$$

Sensitivity and Specificity

- Consider only 1% of population is diseased
- Let's say:
  - $t = 1$, for diseased
  - $t = 0$, for healthy

- What if algorithm always predicts $t = 0$?
  - $Sen =?$
  - $Sen = 0$
  - $Spec =?$
  - $Spec = 1$

- We would like both Sen and $Spec$ to be as high as possible
  - Often, increasing one will decrease the other

- In a disease diagnosis system:
  - We can probably tolerate a decrease in specificity (healthy samples classified as diseased), if it provides an increase in sensitivity (diseased samples classified as healthy)

Performance evaluation

- Let's recall that often classification algorithms provide real-valued output which is then thresholded to assign a classification label

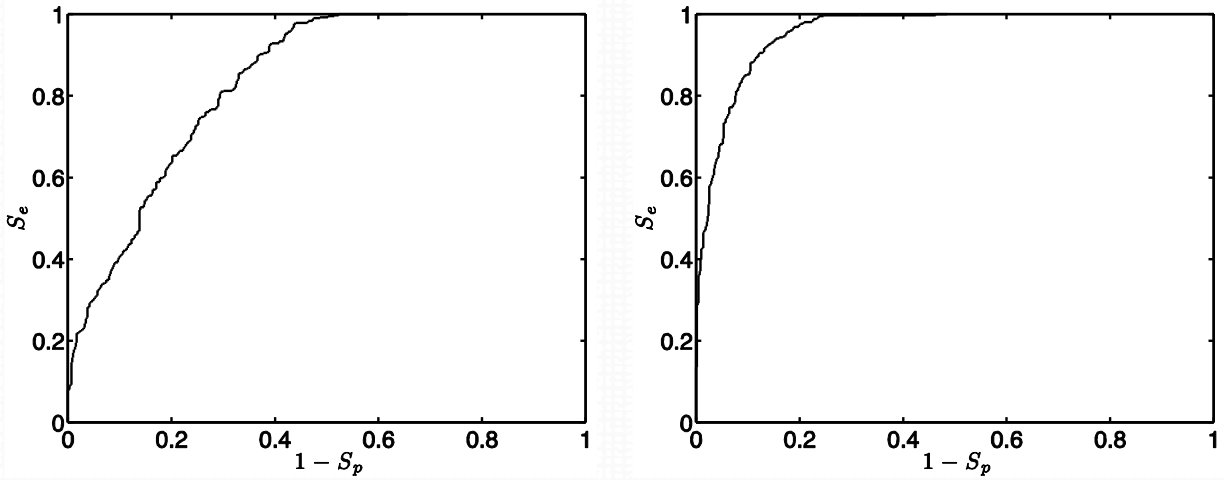
- Bayesian classifier
  - $P(t_{new} = 1 | \boldsymbol{x}_{new}, \boldsymbol{X}, t)$

- SVM
  - $t_{new} = sign(\sum_{n=1}^{N} t_n \alpha_n k(\boldsymbol{x}_n, \boldsymbol{x}_{new}) + b)$

- How about perturbing this threshold value?

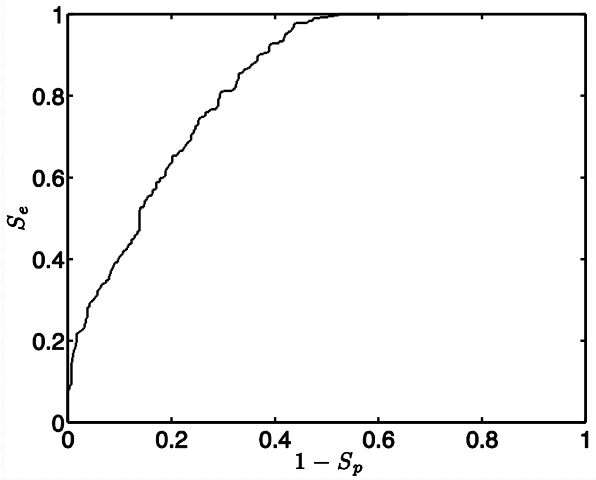
ROC analysis

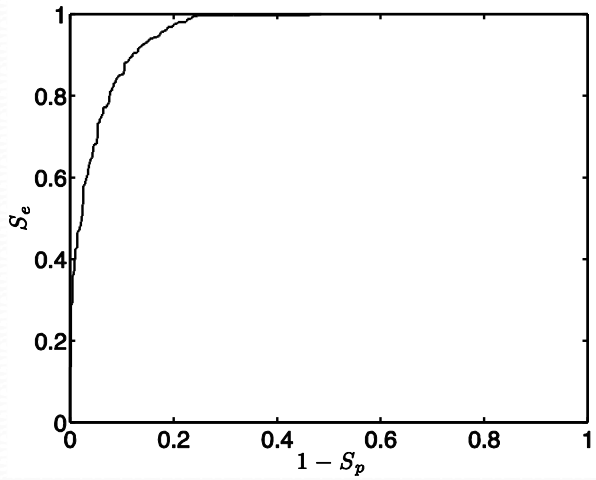
- Threshold is changed to evaluate classifier's performance
- Receiver operating characteristic (ROC) curve
  - Sensitivity (Sen) is plotted against the complementary specificity ($1 - Spec$)
  - Every point on the curve reflects classifier's performance at a particular threshold value



ROC analysis

- We would like both Sen and $Spec$ to be as high as possible
  - i.e. $Sen = 1, Spec = 1, 1 - Spec = 0$
- Bottom left: every sample is classified as healthy (0)
- Top right: every sample is classified as diseased (1)
- Ideal: get the curve to the top left corner
  - Perfect classification: $Sen = 1, Spec = 1$

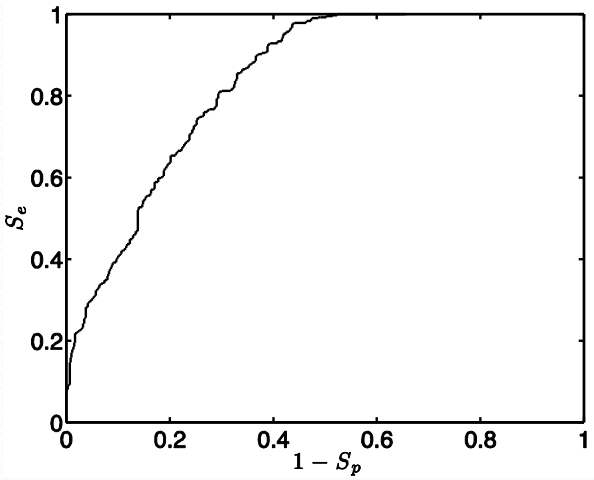




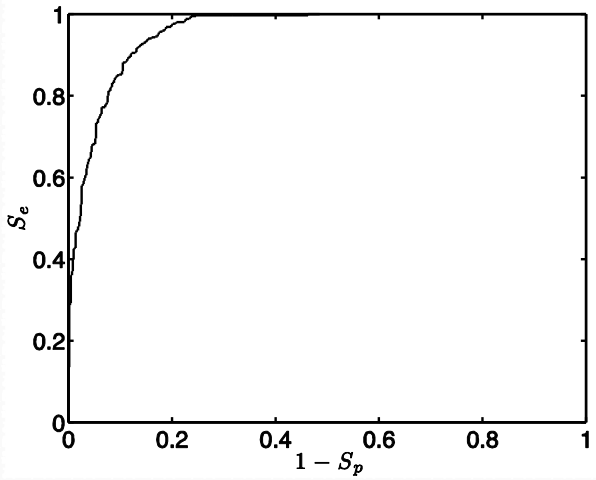
ROC analysis

- Area under curve (AUC)
  - Quantify performance by estimating AUC
  - The higher, the better
- AUC is a better measure than 0/1 loss or classification accuracy
  - Considers class imbalance in data

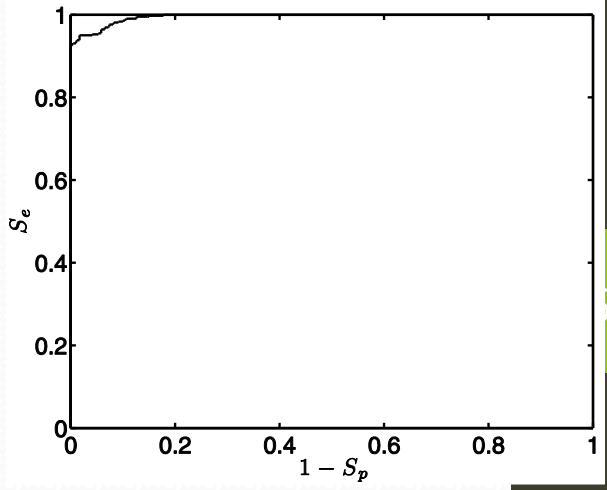
AUC= 0.8348



AUC= 0.9551



AUC= 0.9936



ROC analysis

- Multi-class classification
  - Not naturally suitable
- One-against-all classification
- For example: for a 3 class problem, we will generate 3 ROC analyses
  - Each one looks at the binary classification results for class c against the rest

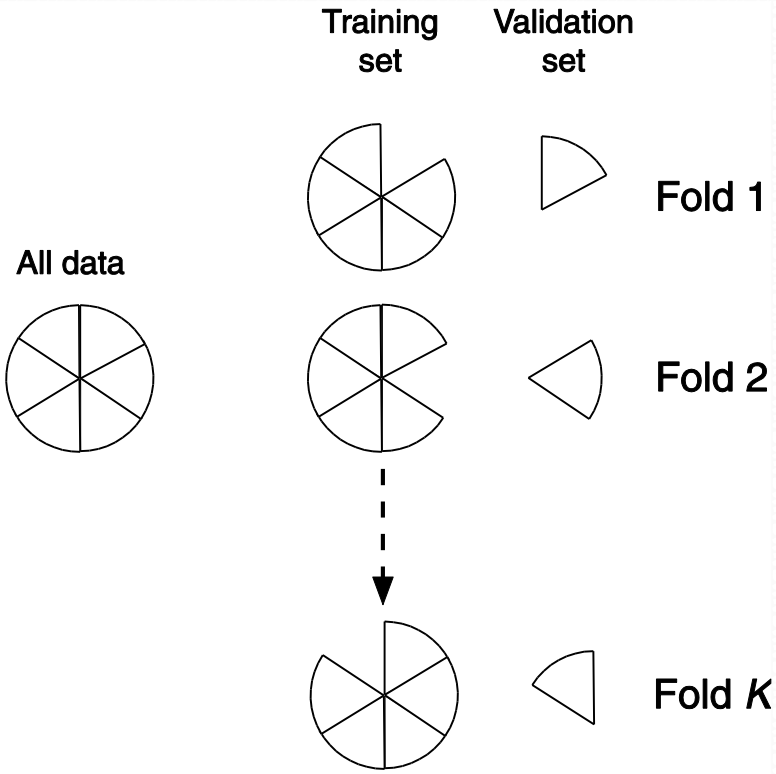
Hold-out validation

- We have earlier discussed the issues of overfitting and generalization
  - Ideally, we want a classifier that can generalize well i.e. on unseen data
  - Where to get unseen data?

- Hold-out validation
  - Partition observed data in to training and testing portions
  - For example: 80-20, or 50-50 split
- Disadvantages
  - Reduce training data
  - Representativeness i.e. validation is biased towards choice of data in validation set, particularly if data is small

Cross-validation

- K-fold cross-validation
  - Hold out K^{th} portion/fold for testing
  - Repeat the classifier training and testing for each fold
- Compute average error from all folds
- Leave-one-out cross-validation (LOOCV)
  - Extreme case of K-fold cross-validation
  - Computationally exhausting



Repeated cross-validation

- The hold-out validation or k-fold cross-validation can be made more reliable by repeating it several times
  - With random selection of training and testing datasets
- Accuracies obtained from various repeats are averaged to indicate overall performance
- Computationally exhaustive

Summary

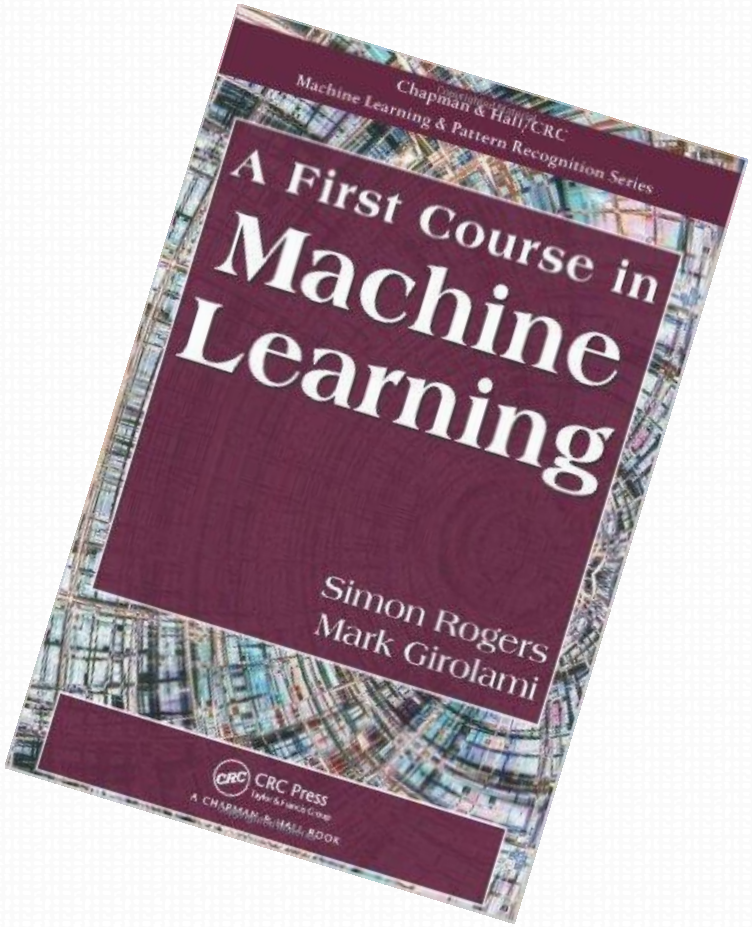
- 0/1 loss
- Classification accuracy

- Confusion matrix
- Sensitivity
- Specificity

- ROC analysis

- Assessing binary-class classification performance
- Assessing multi-class classification performance

- Cross-validation



Author's material
(Simon Rogers)

Thank You