QUEEN'S UNIVERSITY BELFAST
SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

Data Analysis & Visualisation

CSC3062

BEng (CS & SE), MEng (CS & SE), BIT & CIT

Dr Reza Rafiee

Semester 1 2019

Supervised learning

What we need to know about classification

- What is classification?
- What we need as a dataset in classification
- Binary vs. multiclass classification
- Classification models (categories of classifier models)
- How to choose a classification model?
- Support vector machine (SVM) classifier model
- Designing a multiclass SVM model with an example
- How to evaluate the performance of a classifier model?

Tuning the model; grid search and 10-fold cross validation

TUNING:

```
Tuning_model <- tune(svm, Trainingset450k17, label_vector,
scale = F, tolerance = 0.00001, type = "C-classification",
kernel = "radial", probability = T
ranges = list(cost= seq(8, 12, 1), gamma = seq(0.20, 0.25, 0.01)),
tunecontrol= tune.control(sampling = "cross", cross=10), seed=123456)

Plot(Tuning_model, xlime=range(0:15), ylime=range(0:1))

Plot(Tuning_model, xlime=range(0.2:0.25), ylime=range(8:12))
```

The darkest shades of blue indicating the best (see the two plots).
Narrowing in on the darkest blue range and performing further tuning.

2) TRAINING:

```
Radial_model <- svm(Trainingset450k17, label_vector, scale = F,
tolerance = 0.00001, type = "C-classification",
kernel = "radial",
cost = 10, gamma = 0.22,
probability = T, seed = 123456)
```

3) TESTING (PREDICTION):

```
Radial_model <- predict(object= Radial_model, newdata = seq_test_BEM_97, probability=T)
```

Three key steps

1) Tuning

Choose a hyperplane; try linear or nonlinear (polynomial or RBF kernels) and find it's parameters

2) Training

Train the classifier based on the identified

parameters of the hyperplane

3) Testing

Test the trained classifier by giving it some new samples (without subgroups): seq_test_BEM_97

What is resampling technique?

If you use the entire training data to select the "optimal" classifier, then there would be a fundamental problem.

The final model will normally **overfit** the training data: it will not be able to generalise to new data.

The error rate estimate will be overly optimistic (lower than the true error rate)

Split dataset into two groups

Training set: used to train the classifier

Test set: used to estimate the error rate of the trained classifier

K-fold cross-validation (CV)

Cross validation and bootstrapping are resampling methods

Question: why do we need resampling method?

A limited number of good samples (limited data)

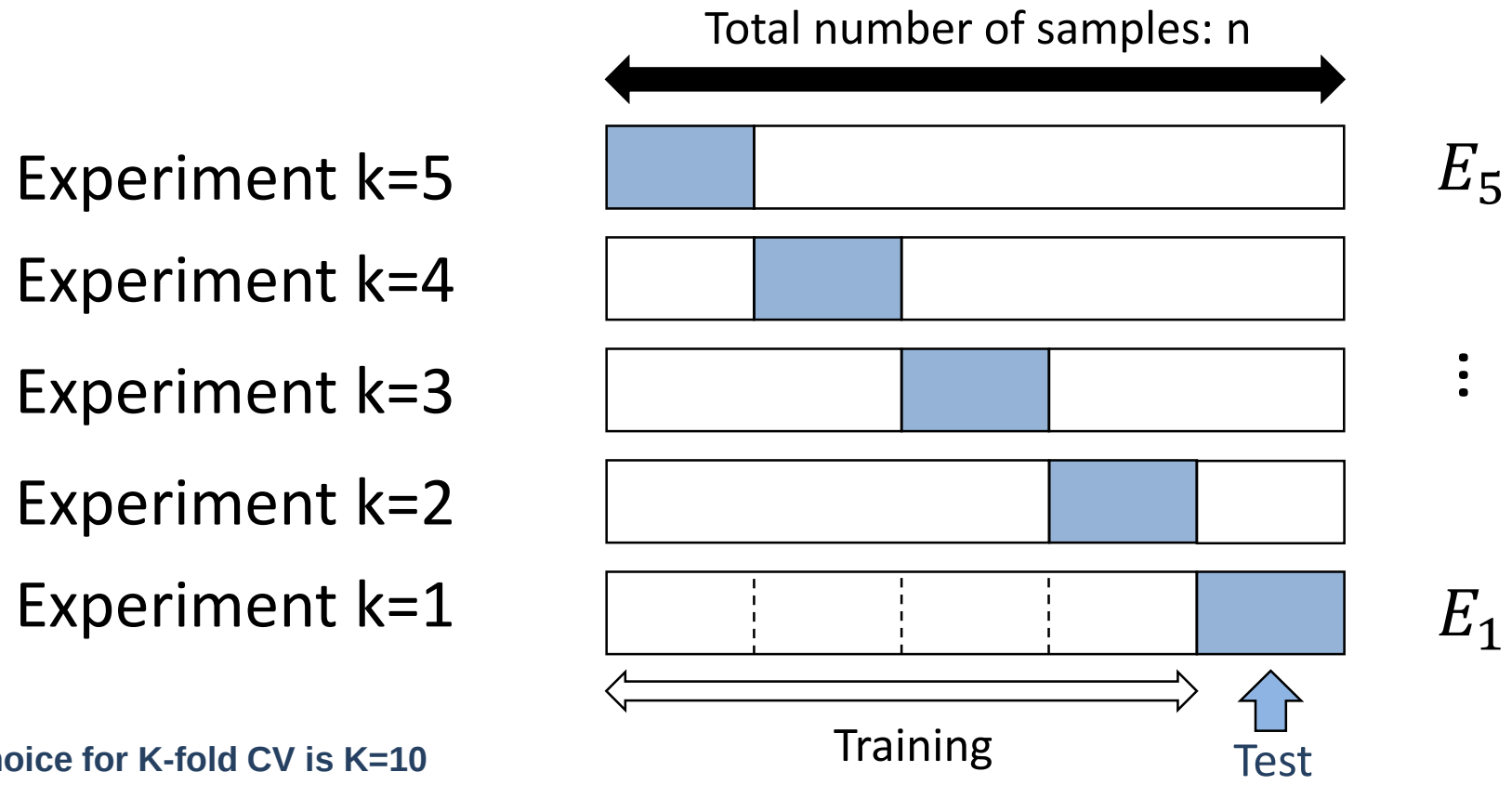
Collection of data is expensive

K-fold cross-validation (CV)

Create a K-fold partition of a dataset

For each of K experiments, use K-1 folds for training and a different fold for testing

This procedure is illustrated in the following figure for K=5



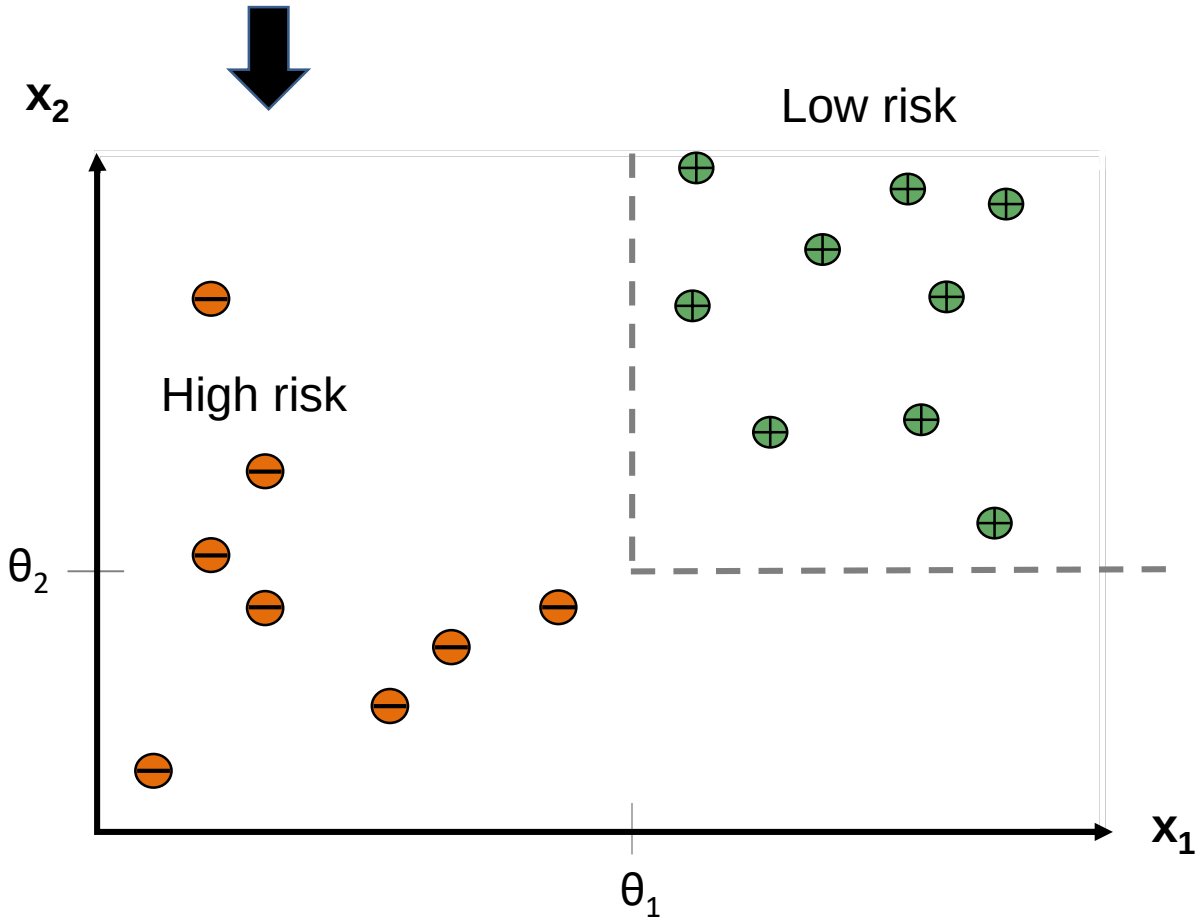
$$E = \frac{1}{K}\sum_{i=1}^{K} E_i$$

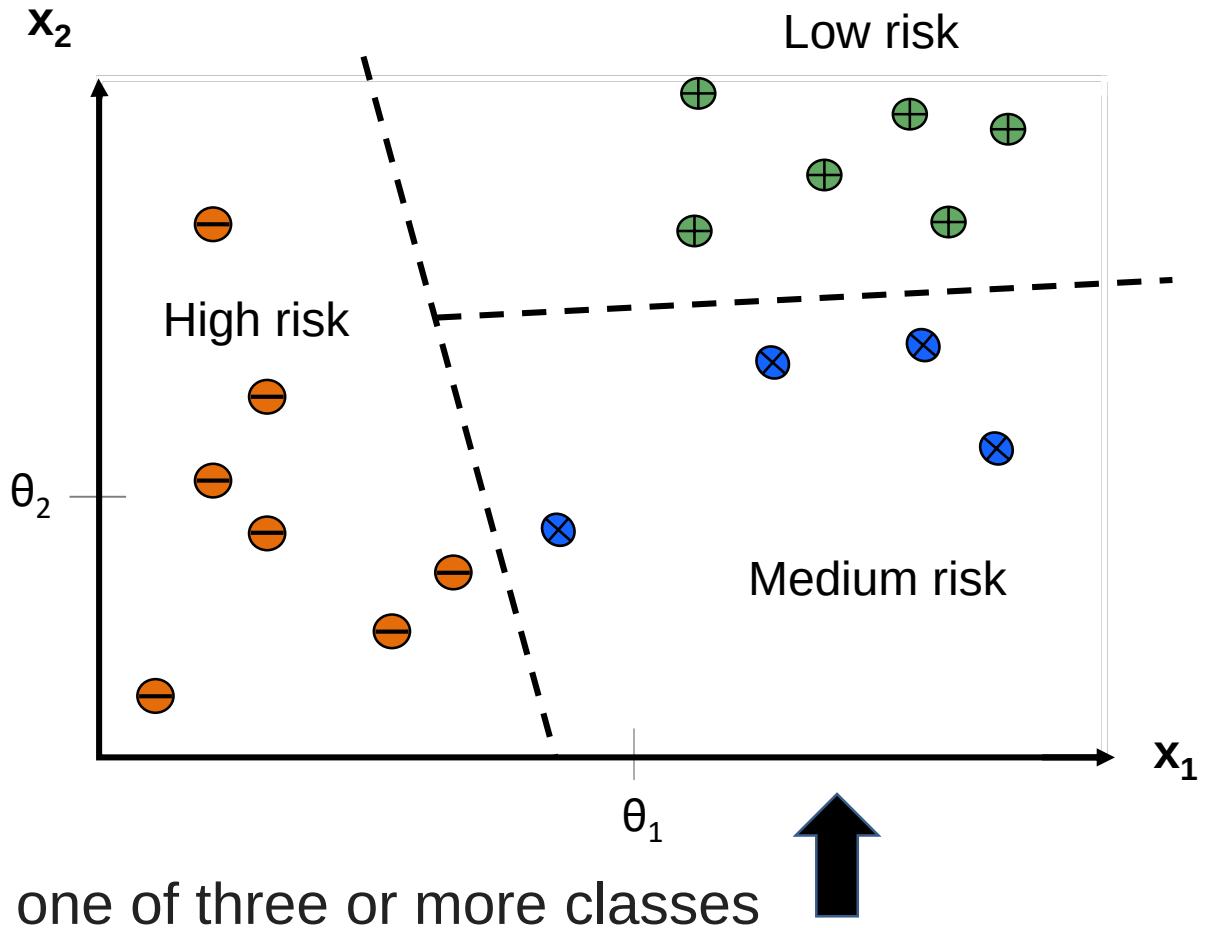
Average error

A common choice for K-fold CV is K=10

Binary vs. multiclass classification

Binary classifier classifies data points into one of two classes

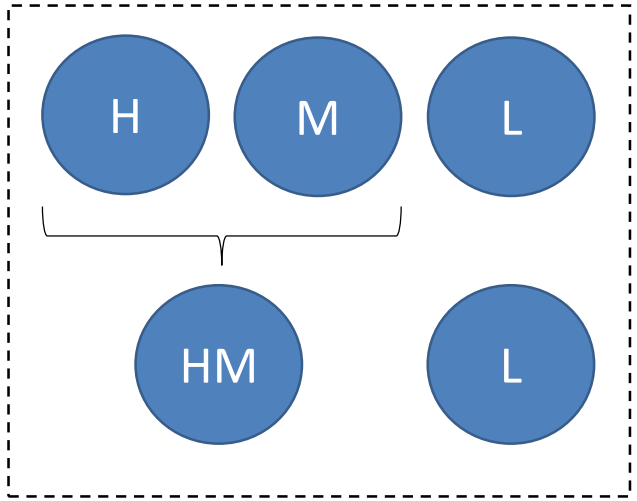


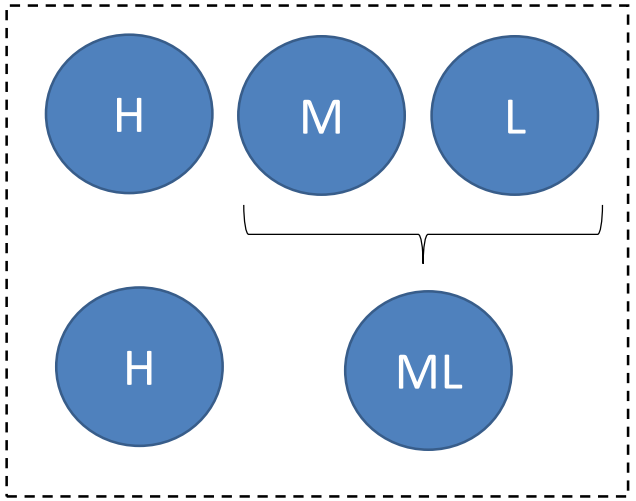


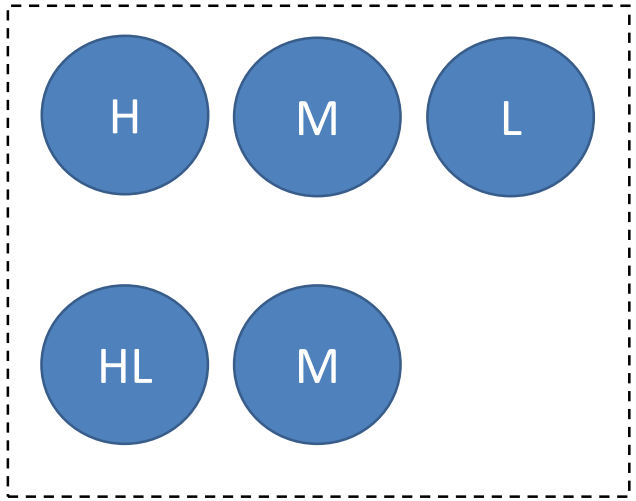
Multiclass classifier: classifies data points into one of three or more classes

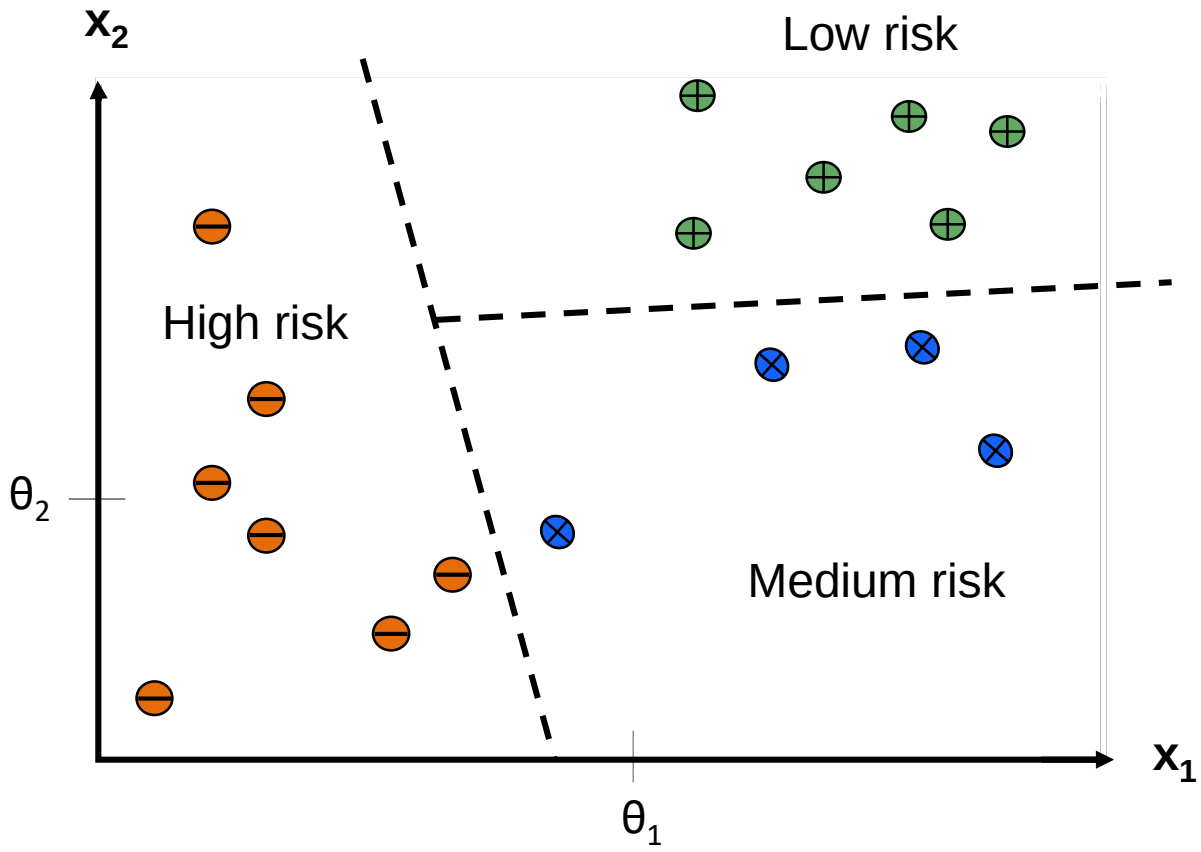
Multiclass to binary classification

High risk: H Medium risk: M Low risk: L







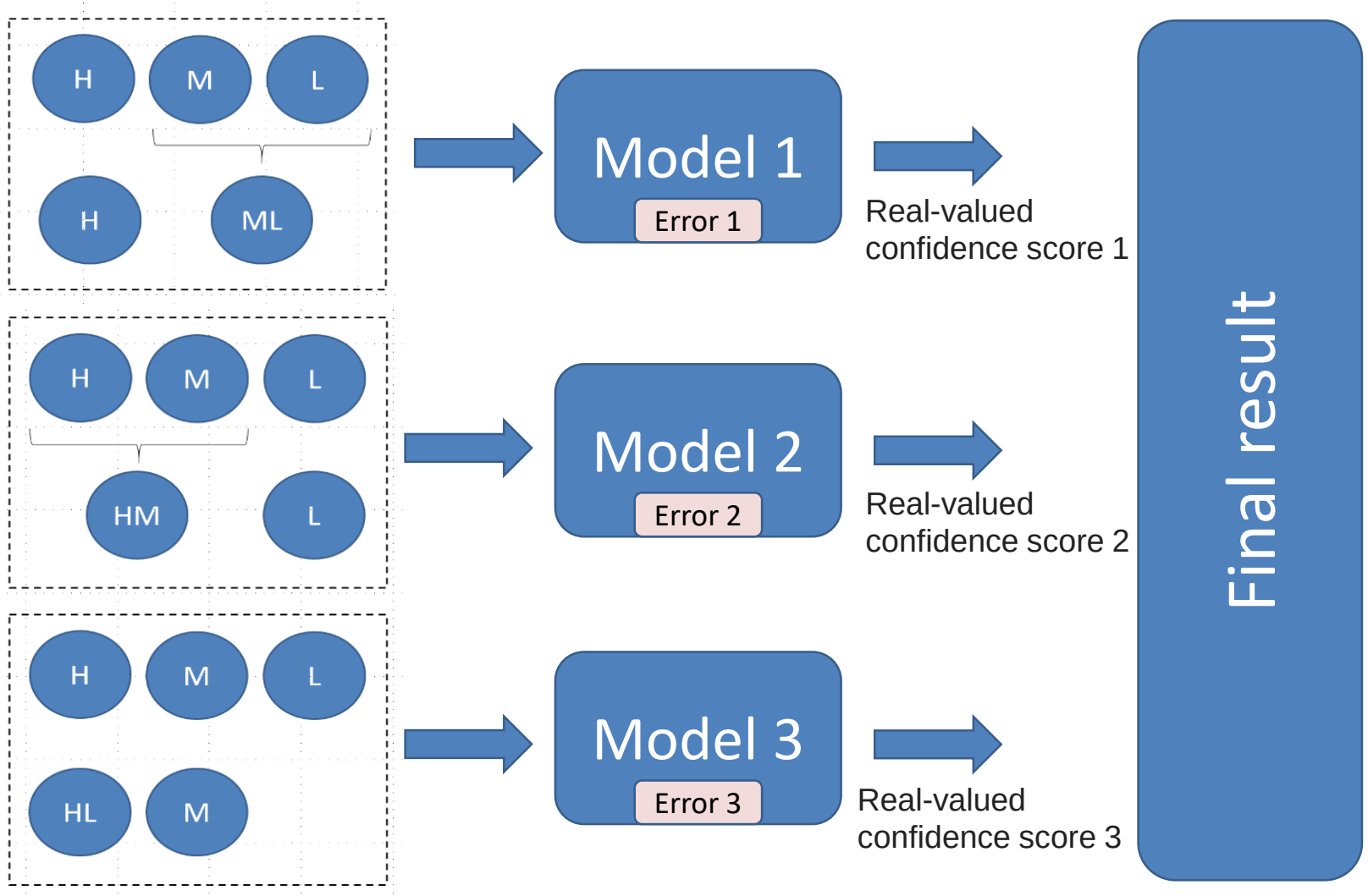


One vs. rest (all) approach

One vs. one approach

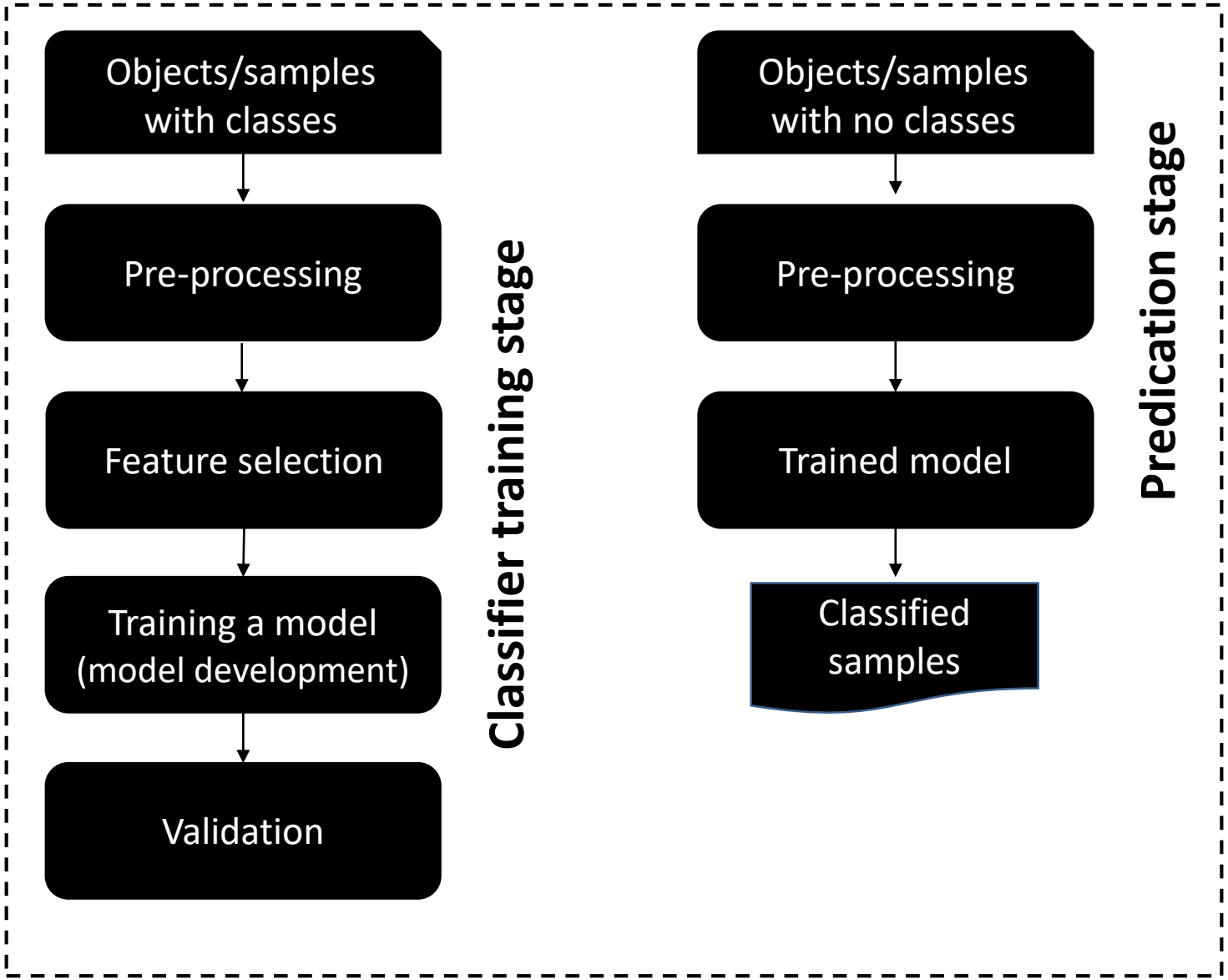
Multiclass to binary classification

Training stage



Training a single classifier per class

Classification



Evaluate classification performance

- Confusion Matrix
  - True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN)
- Accuracy
- Precision
- Recall or Sensitivity
- Specificity
- F1 Score

https://en.wikipedia.org/wiki/Confusion_matrix

Confusion matrix (error matrix) - multiclass

Describe the performance of a multiclass classification model using a confusion matrix

Reference subgroup (actual subgroup)

Predicted by a classifier	Group 1	Group 2	Group 3	Group 4
Group 1	**16**	0	0	0
Group 2	0	**31**	**1**	0
Group 3	0	0	**19**	0
Group 4	0	0	0	**45**

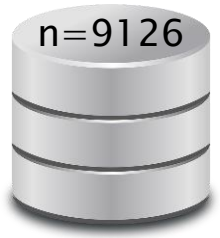
Reference

	Group 1	Group 2
Group 1	**TP**	FP
Group 2	FN	**TN**

True positive: TP
False positive: FP
True negative: TN
False negative: FN

Classifier model based on 9126 samples and 440 genes

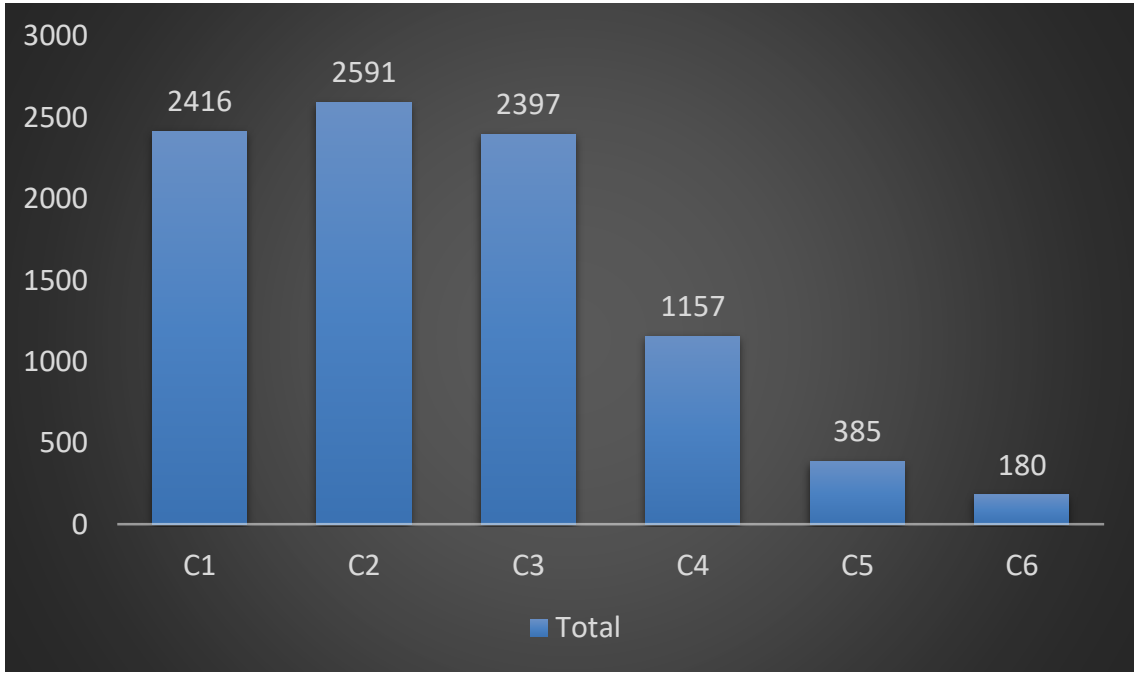
Original dataset



Training data (with subgroup)

6 immune subgroups
440 genes (5 signatures)

Immune subtype	Number of samples in each immune subtype
C1	2416
C2	2591
C3	2397
C4	1157
C5	385
C6	180

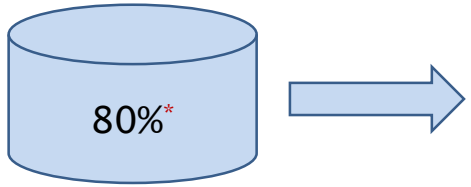


QUEEN'S UNIVERSITY BELFAST
SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

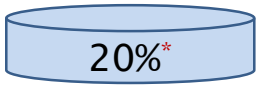
Optimising the parameters of a nonlinear SVM classifier

RNA-seq, 6 immune subgroups
440 genes (5 signatures)

- MultipleImputationModelling()
- Dataset_Spilitter() including
- TCGA_Cancer_Immune_Subtypes()
- TCGA_Gene_Signature()
- Index_of_Sample_When_Spilit()



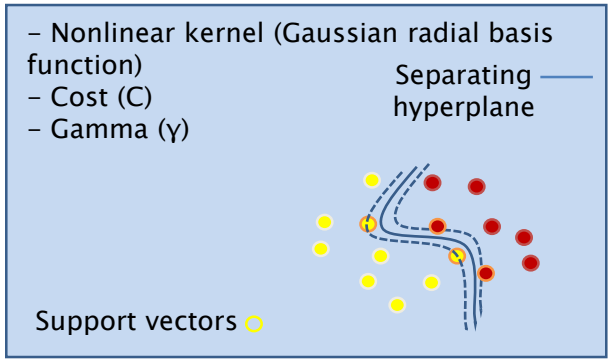
Training/validation dataset



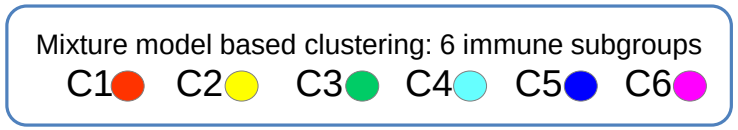
Test dataset
(Unobserved data)

*80% - 20% of each class

Multi-class Support vector machine (SVM) classifier



- One-against-one approach
- Tried linear, polynomial and RBF kernels and RBF kernels performed the best
- Tuning (optimising) C and γ using a grid search and 10-fold cross validation technique
  - Building models for multiple combinations of parameter values and selecting the best.

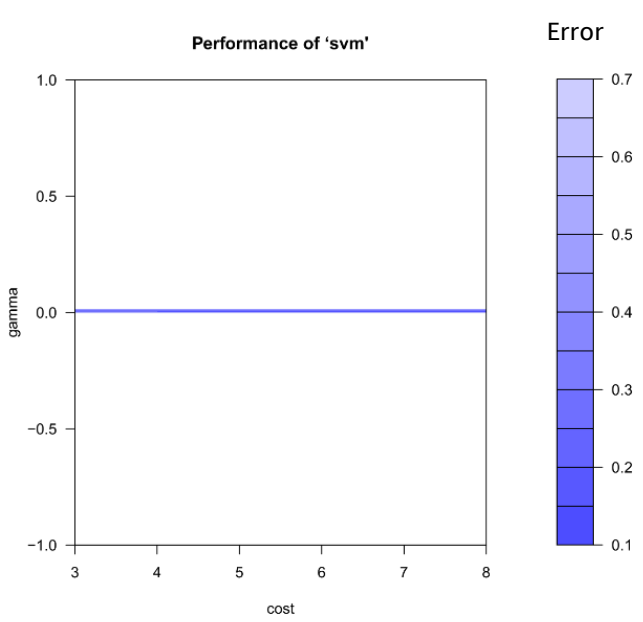


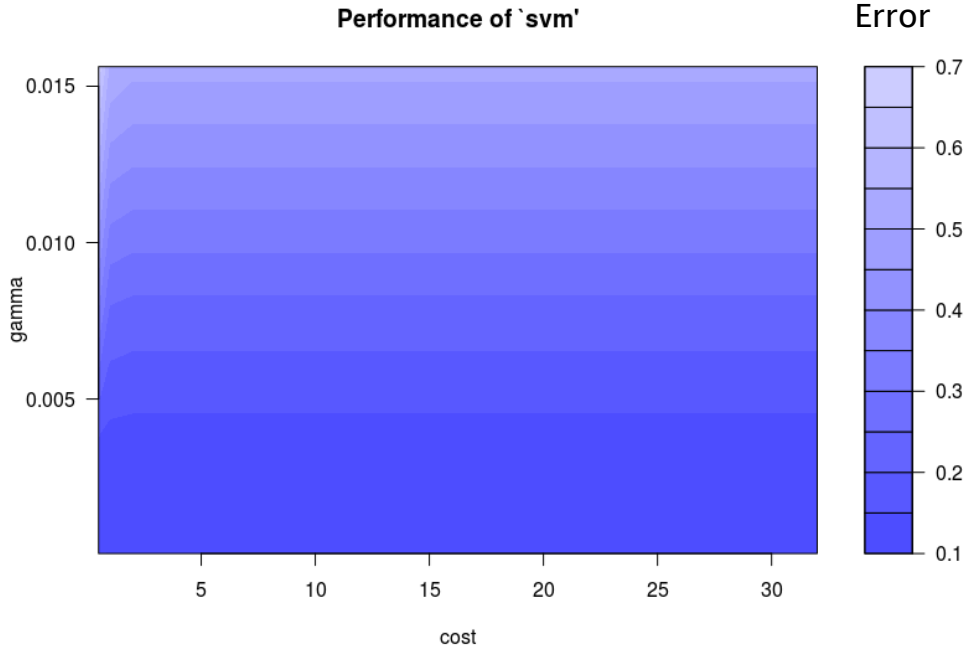
Based on the "Immune Landscape of Cancer" paper

- When C is large, the margin is wide, and there are many support vectors.
- When C is small, we seek narrow margins that are rarely violated (low bias, high variance).
- C controls the bias-variance trade-off.
- γ controls the standard deviation of the Gaussian function.

QUEEN'S UNIVERSITY BELFAST
SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

Grid search cross-validated training (n=7300)

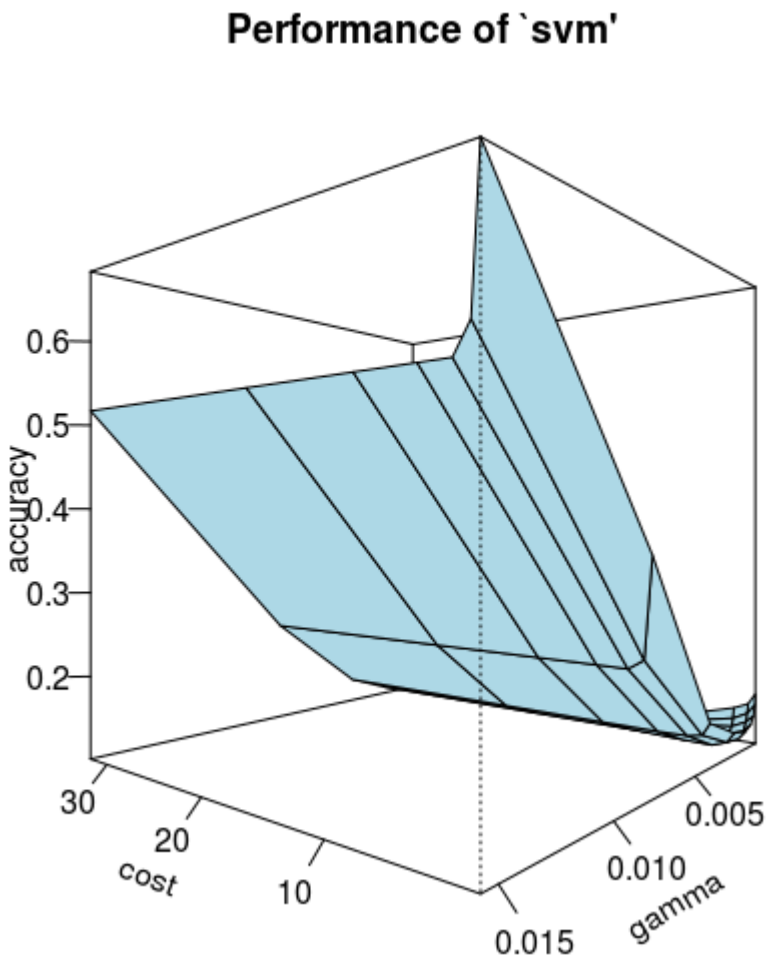


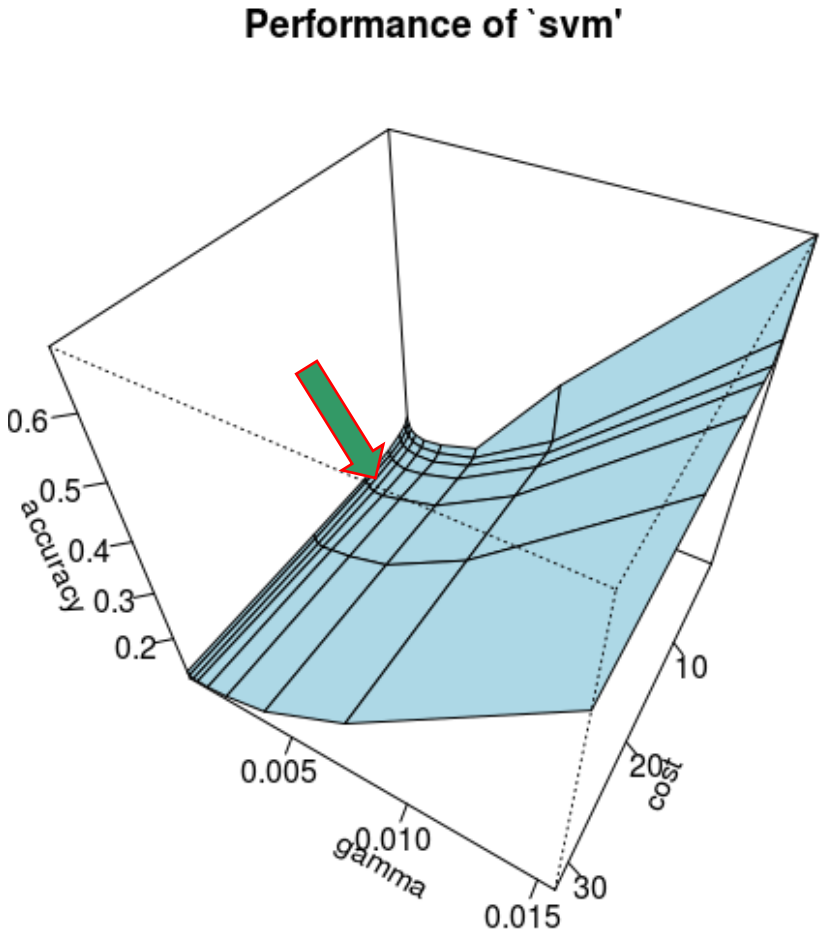


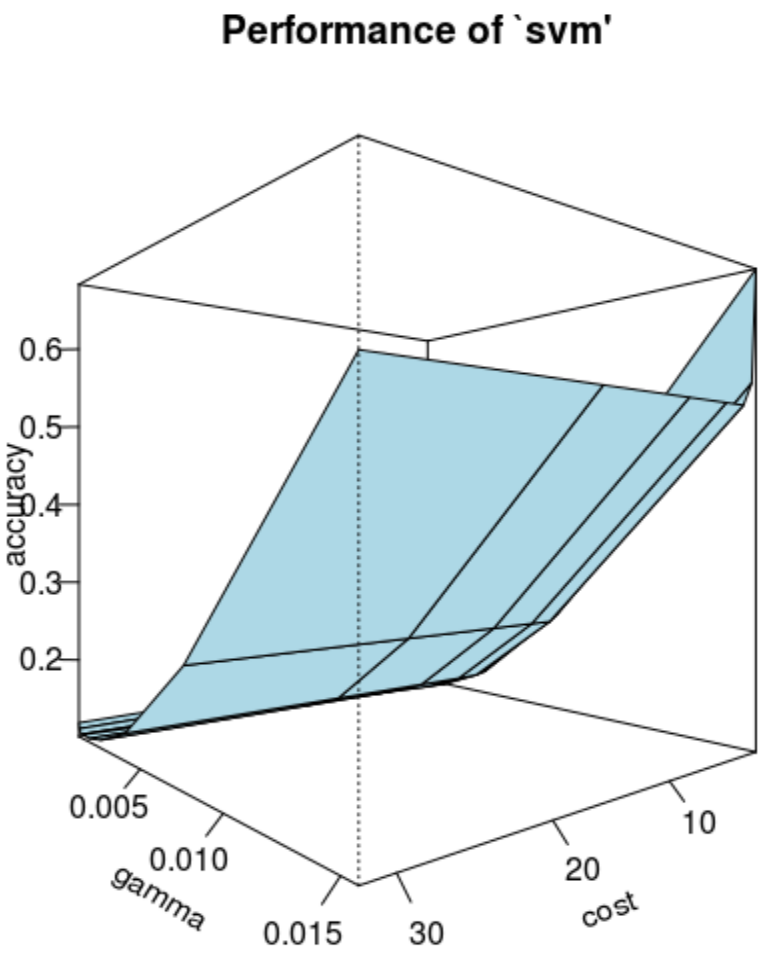
- Initial range: cost = 2^(-1:5), gamma=2^(-14:-6)
- 10-fold cross validation
- Computational time (on Kelvin Clusters) ~144 hours (intensive)
- Best parameters : c=8, γ= 0.000977
- Best performance (cross-validation accuracy): 89.85%
- Misclassification error (MSE) used for assessing

#	cost	gamma	error	dispersion
1	0.5	6.10E-05	0.16434	0.016693
2	1	6.10E-05	0.150643	0.014414
3	2	6.10E-05	0.142974	0.015261
4	4	6.10E-05	0.133251	0.016698
5	8	6.10E-05	0.126266	0.016624
6	16	6.10E-05	0.122569	0.016242
7	32	6.10E-05	0.119145	0.015617
8	0.5	0.000122	0.15037	0.015097
9	1	0.000122	0.138866	0.015577
10	2	0.000122	0.129826	0.017489
11	4	0.000122	0.122568	0.016091
12	8	0.000122	0.120104	0.017211
13	16	0.000122	0.115037	0.013512
14	32	0.000122	0.112026	0.011304
15	0.5	0.000244	0.137496	0.016093
16	1	0.000244	0.128731	0.017364
17	2	0.000244	0.120241	0.015275
18	4	0.000244	0.115859	0.01703
19	8	0.000244	0.109422	0.013913
20	16	0.000244	0.105452	0.011629
21	32	0.000244	0.105588	0.013308
22	0.5	0.000488	0.128595	0.018547
23	1	0.000488	0.118325	0.017501
24	2	0.000488	0.112984	0.017151
25	4	0.000488	0.108189	0.015978
26	8	0.000488	0.102576	0.012763
27	16	0.000488	0.102165	0.013166
28	32	0.000488	0.106958	0.013633
29	0.5	0.000977	0.122981	0.018059
30	1	0.000977	0.112435	0.017506
31	2	0.000977	0.108327	0.016439
32	4	0.000977	0.102849	0.013686
33	**8**	**0.000977**	**0.10148**	**0.013272**
34	16	0.000977	0.105725	0.013511
35	32	0.000977	0.105725	0.014026
36	0.5	0.001953	0.127913	0.01585
37	1	0.001953	0.115313	0.015628
38	2	0.001953	0.109012	0.013842
39	4	0.001953	0.109286	0.012823
40	8	0.001953	0.111204	0.012277

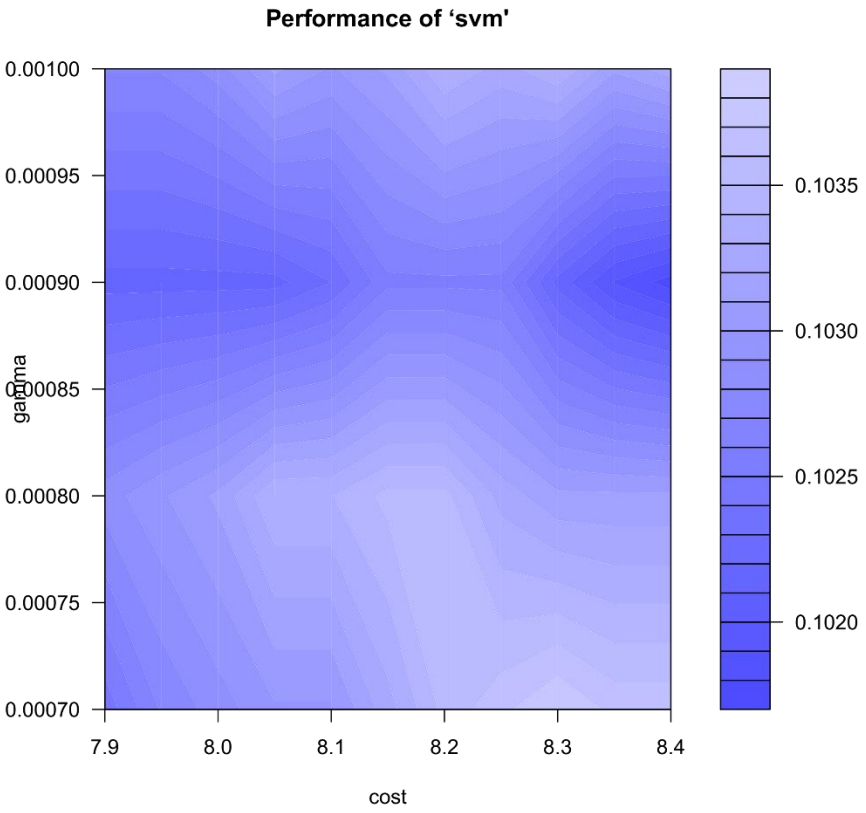
Hyperplane (3 angles)







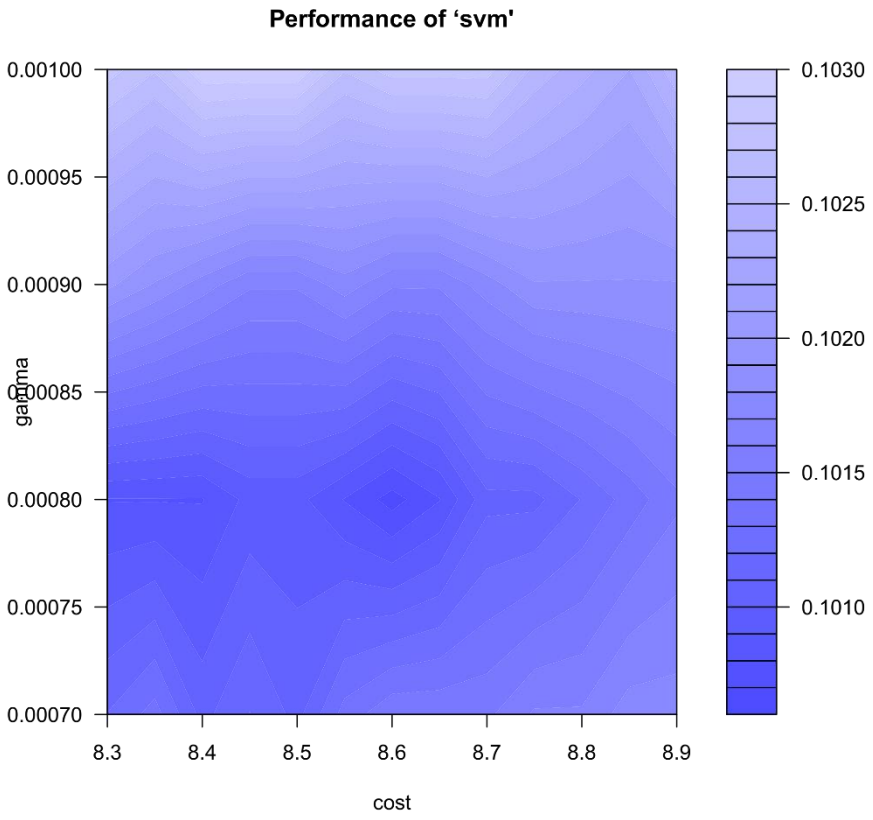
Refine the model, ver1 (not improved)



- New range: cost =seq(7.9,8.4,0.05), gamma=seq(0.0007,0.001,0.0001)
- 10-fold cross validation
- Computational time: ~72 hours
- Best parameters : c=8.4, γ= 0.0009
- Best performance (cross-validation accuracy): 89.82%

#	cost	gamma	error	dispersion
1	7.9	7.00E-04	0.102437	0.015038
2	7.95	7.00E-04	0.102711	0.014984
3	8	7.00E-04	0.102848	0.014766
4	8.05	7.00E-04	0.102985	0.014786
5	8.1	7.00E-04	0.102985	0.014601
6	8.15	7.00E-04	0.103259	0.014806
7	8.2	7.00E-04	0.103533	0.014565
8	8.25	7.00E-04	0.10367	0.014929
9	8.3	7.00E-04	0.103807	0.014589
10	8.35	7.00E-04	0.10367	0.014563
11	8.4	7.00E-04	0.10367	0.014563
12	7.9	8.00E-04	0.102847	0.015517
13	7.95	8.00E-04	0.102984	0.015319
14	8	8.00E-04	0.103121	0.015363
15	8.05	8.00E-04	0.103395	0.015162
16	8.1	8.00E-04	0.103395	0.015162
17	8.15	8.00E-04	0.103532	0.015106
18	8.2	8.00E-04	0.103532	0.015285
19	8.25	8.00E-04	0.103259	0.015039
20	8.3	8.00E-04	0.103122	0.015105
21	8.35	8.00E-04	0.103122	0.015133
22	8.4	8.00E-04	0.103122	0.015133
23	7.9	9.00E-04	0.102163	0.015314
24	7.95	9.00E-04	0.102163	0.015314
25	8	9.00E-04	0.102163	0.015314
26	8.05	9.00E-04	0.102163	0.015314
27	8.1	9.00E-04	0.1023	0.015339
28	8.15	9.00E-04	0.102574	0.015209
29	8.2	9.00E-04	0.102574	0.015087
30	8.25	9.00E-04	0.102574	0.015087
31	8.3	9.00E-04	0.102163	0.015071
32	8.35	9.00E-04	0.10189	0.014722
33	**8.4**	**9.00E-04**	**0.101753**	**0.014649**
34	7.9	0.001	0.102711	0.01497
35	7.95	0.001	0.102711	0.01497
36	8	0.001	0.102848	0.015143
37	8.05	0.001	0.103122	0.015165
38	8.1	0.001	0.102985	0.015148
39	8.15	0.001	0.103122	0.014998
40	8.2	0.001	0.103396	0.015181
41	8.25	0.001	0.103259	0.015015
42	8.3	0.001	0.103396	0.015152
43	8.35	0.001	0.103122	0.015093
44	8.4	0.001	0.103259	0.015286

Refine the model, ver2 (slightly improved)



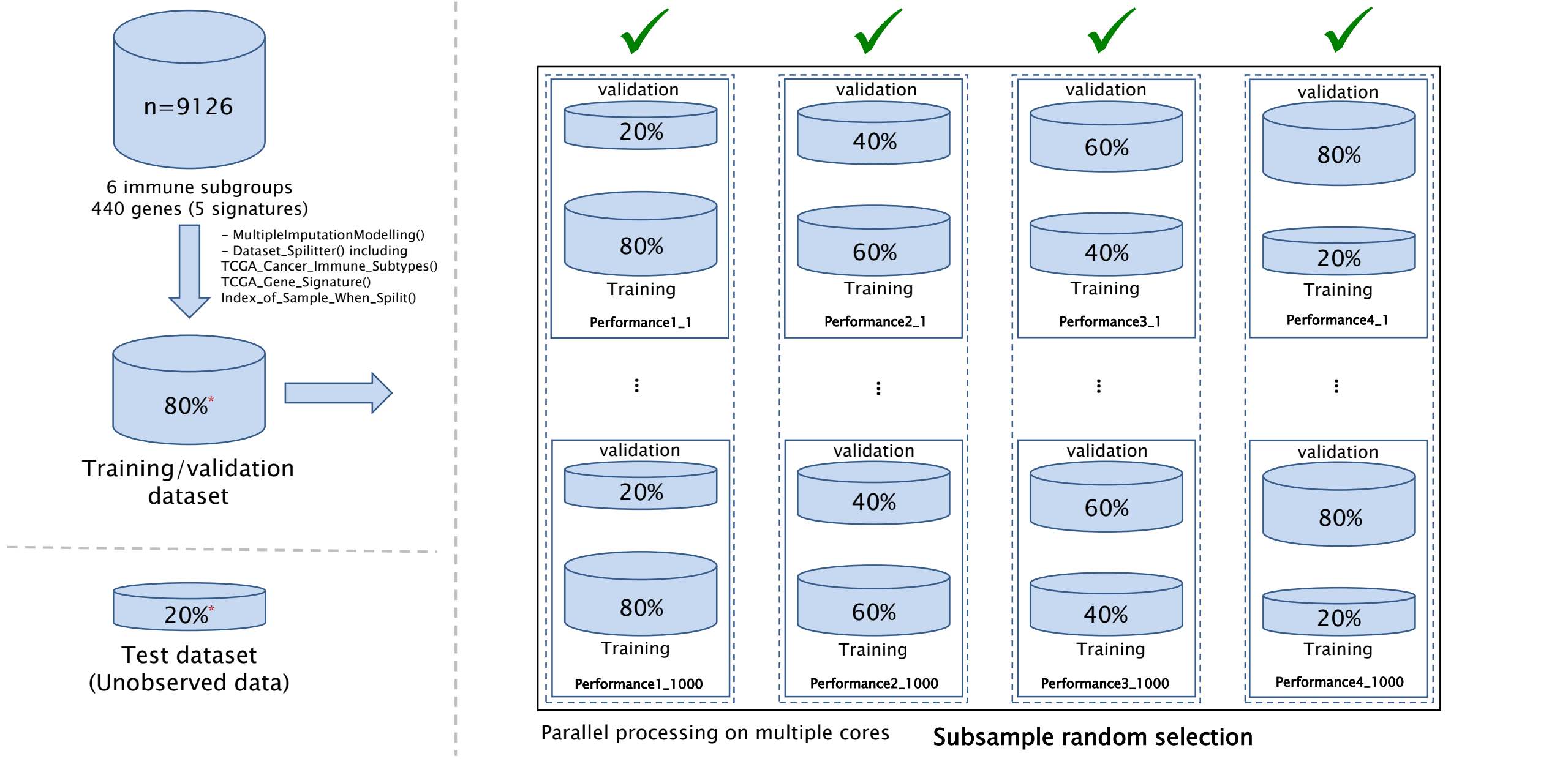
	#	cost	gamma	error	dispersion
➔	20	8.6	8.00E-04	0.100657384	0.007941988
	16	8.4	8.00E-04	0.100794183	0.00804287
	14	8.3	8.00E-04	0.100794371	0.007888516
	15	8.35	8.00E-04	0.100794371	0.007888516
	19	8.55	8.00E-04	0.100794371	0.007941203
	21	8.65	8.00E-04	0.100794371	0.00780882
	17	8.45	8.00E-04	0.10093117	0.007961321
	18	8.5	8.00E-04	0.10093117	0.007961321
	3	8.4	7.00E-04	0.101067219	0.008199534
	5	8.5	7.00E-04	0.101067219	0.008199534
	22	8.7	8.00E-04	0.101068343	0.007499372
	23	8.75	8.00E-04	0.101068343	0.007499372
	1	8.3	7.00E-04	0.101204205	0.007932533
	4	8.45	7.00E-04	0.101204205	0.008036983
	24	8.8	8.00E-04	0.10120533	0.007682581
	2	8.35	7.00E-04	0.101341191	0.007921247
	6	8.55	7.00E-04	0.101341191	0.007999823
	25	8.85	8.00E-04	0.101342316	0.007670905
	7	8.6	7.00E-04	0.101478178	0.008192227
	8	8.65	7.00E-04	0.101478178	0.008418173
	9	8.7	7.00E-04	0.101478178	0.008418173
	26	8.9	8.00E-04	0.101479115	0.007680895
	10	8.75	7.00E-04	0.101615164	0.008402578

Sorted based on error (only 23 out of 52 rows illustrated)

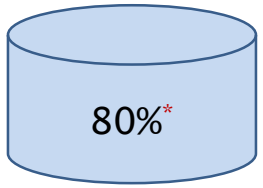
- New range: cost =seq(8.3,8.9,0.05), gamma=seq(0.0007,0.001,0.0001)
- 10-fold cross validation
- Computational time: ~ 72 hours
- Best parameters : c=8.6, γ= 0.0008
- Best performance (cross-validation accuracy): 89.93%

QUEEN'S UNIVERSITY BELFAST
SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

Performance evaluation of training models



Confusion matrix and overall statistics



Training/validation dataset

Training (n=7302), validation (n=7302)
Accuracy : 0.9969
95% CI : (0.9953,0.998)
Kappa: 0.9959

n=23 discordant

Classifier ver1.1.0	Reference (model based clustering) C1	C2	C3	C4	C5	C6	NC	Total
C1	1926	4	0	0	0	3		1933
C2	2	2069	1	0	0	2		2074
C3	4	0	1916	0	0	5		1925
C4	0	0	0	926	0	0		926
C5	0	0	0	0	308	0		308
C6	1	0	1	0	0	134		136
NC								
Total	1933	2073	1918	926	308	144		7302

	Sensitivity	Specificity	Pos-Pred Value	Neg-Pred Value	Precision	Recall	F1	Prevalence	Detection Rate	Detection Prevalence	Balanced Accuracy
Class: C1	0.996378686	0.998696219	0.996378686	0.998696219	0.996378686	0.996378686	0.996378686	0.264721994	0.263763353	0.264721994	0.997537453
Class: C2	0.998070429	0.999043794	0.9975892	0.999234889	0.9975892	0.998070429	0.997829756	0.283894823	0.283347028	0.284031772	0.998557112
Class: C3	0.998957247	0.99832838	0.995324675	0.999628045	0.995324675	0.998957247	0.997137653	0.262667762	0.262393865	0.263626404	0.998642814
Class: C4	1	1	1	1	1	1	1	0.126814571	0.126814571	0.126814571	1
Class: C5	1	1	1	1	1	1	1	0.042180225	0.042180225	0.042180225	1
Class: C6	0.930555556	0.999720592	0.985294118	0.998604521	0.985294118	0.930555556	0.957142857	0.019720624	0.018351137	0.018625034	0.965138074

Overall Statistics for the training dataset (n=7302)

QUEEN'S UNIVERSITY BELFAST
SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

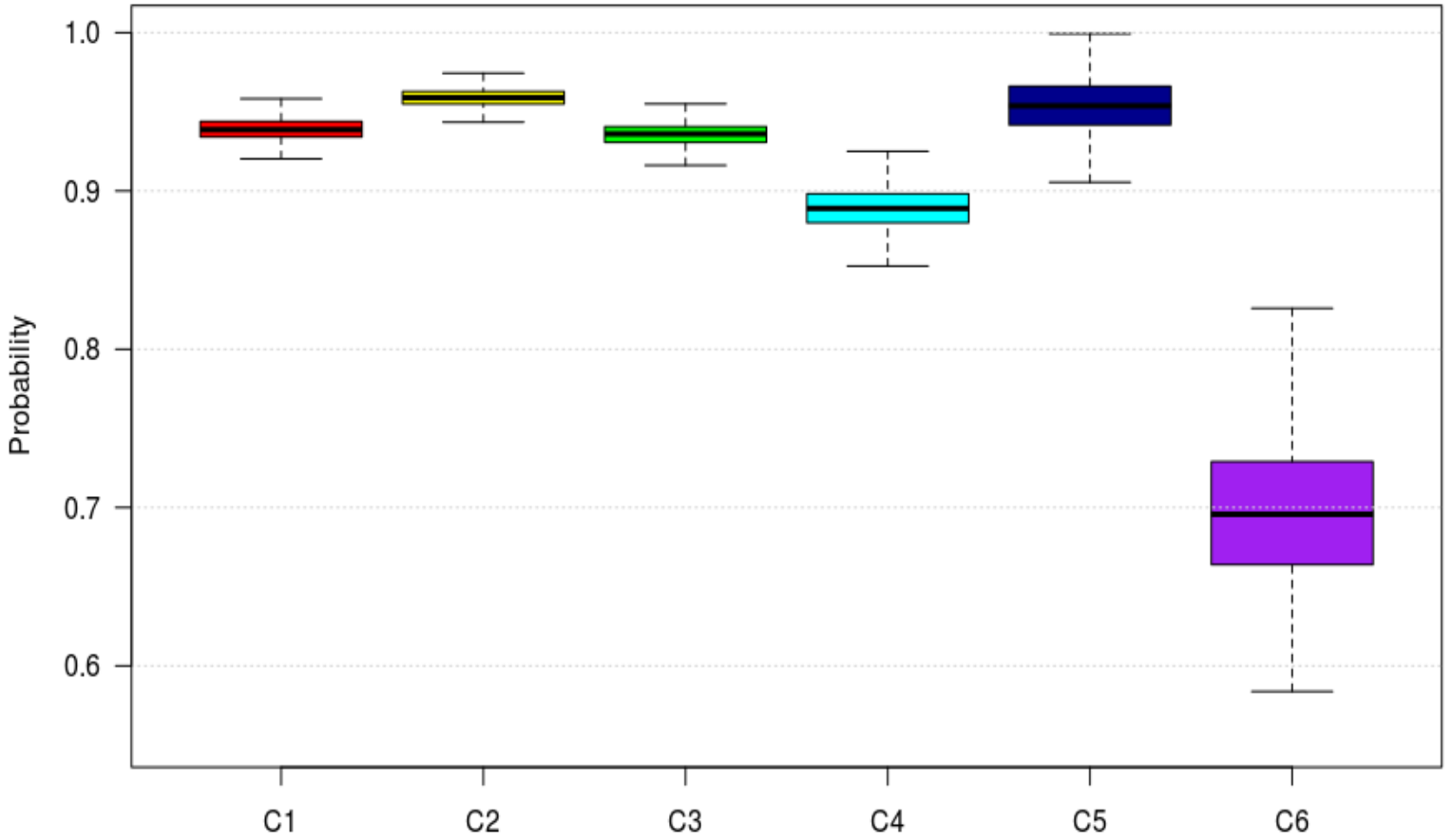
Overall statistics when using the training set for prediction – discordant samples

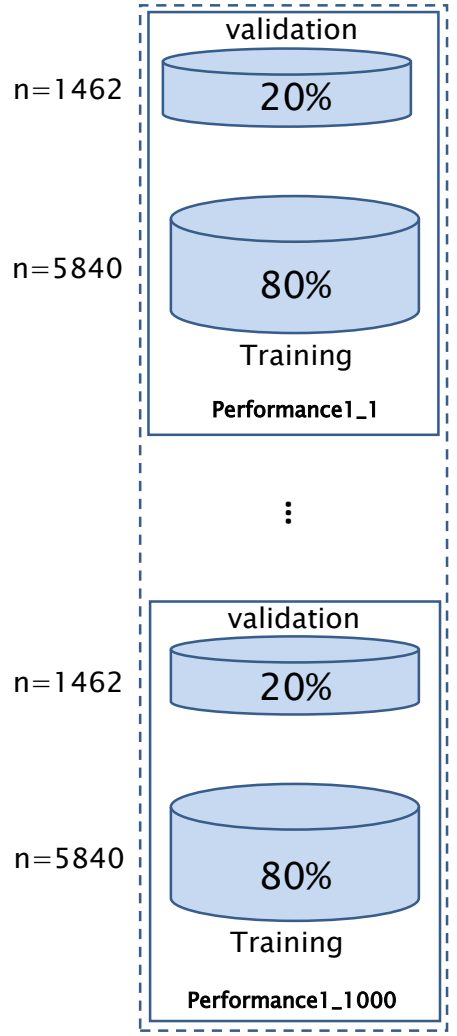
Training set & validation: n=7302; mismatched percentage: 0.3%

	Row_index	Sample_id	Cancer_type	Ref_immune_group (using unsupervised clustering)	Predicted	C1	C2	C3	C4	C5	C6	Problematic Samples*
1	485	TCGA.AA.3514	COAD	C1	C2	0.037	**0.96**	3.60E-05	0.00064	0.00016	5.00E-05	Yes
2	491	TCGA.AA.3525	COAD	C1	C2	0.15	**0.85**	0.0022	0.0021	0.00028	0.0028	Yes
3	716	TCGA.AR.A1AM	BRCA	C1	C3	0.25	0.087	**0.61**	0.0037	0.0027	0.04	No
4	852	TCGA.BG.A2AD	UCEC	C1	C3	0.081	0.00051	**0.91**	0.0028	8.30E-05	0.0017	Yes
5	1506	TCGA.G9.6365	PRAD	C1	C3	0.32	0.00027	**0.67**	0.0013	6.40E-05	0.00032	No
6	1555	TCGA.HZ.A49I	PAAD	C1	C6	0.33	0.067	0.012	0.0027	0.0017	**0.59**	No
7	1880	TCGA.XD.AAUG	PAAD	C1	C3	0.083	0.0012	**0.87**	0.0015	0.0011	0.047	Yes
8	1978	TCGA.21.5782	LUSC	C2	C1	**0.95**	0.055	9.60E-07	6.40E-07	3.90E-05	5.70E-05	Yes
9	2432	TCGA.A6.2671	COAD	C2	C1	**0.99**	0.0017	0.005	0.0014	0.00015	0.00038	Yes
10	2477	TCGA.AA.3522	COAD	C2	C1	**0.91**	0.014	0.0087	0.058	0.0016	0.0045	Yes
11	2756	TCGA.BK.A139	UCEC	C2	C1	**0.99**	0.0088	2.20E-05	0.00019	4.40E-05	3.60E-05	Yes
12	4082	TCGA.55.6970	LUAD	C3	C2	0.0035	**0.75**	0.23	0.006	0.00018	0.016	No
13	5498	TCGA.J2.8192	LUAD	C3	C6	0.0092	0.0011	0.47	0.0023	0.00075	**0.52**	No
14	7163	TCGA.22.1005	LUSC	C6	C2	0.00064	**0.64**	0.048	0.00023	0.00034	0.32	No
15	7169	TCGA.38.7271	LUAD	C6	C3	0.00014	0.002	**0.87**	0.00017	0.0013	0.12	Yes
16	7170	TCGA.3A.A9I7	PAAD	C6	C1	**0.37**	0.088	0.25	0.036	0.028	0.23	No
17	7196	TCGA.75.7030	LUAD	C6	C3	0.0064	4.50E-05	**0.69**	0.00019	0.00044	0.3	No
18	7202	TCGA.98.A53D	LUSC	C6	C3	0.0013	6.40E-05	**0.71**	0.00036	0.00078	0.29	No
19	7206	TCGA.A7.A0DB	BRCA	C6	C3	0.048	0.0016	**0.47**	0.056	0.0018	0.42	No
20	7215	TCGA.AR.A0TT	BRCA	C6	C2	0.0022	**0.62**	0.015	0.00018	0.00047	0.36	No
21	7219	TCGA.AR.A5QM	BRCA	C6	C1	**0.5**	0.024	0.017	0.00066	0.00023	0.46	No
22	7272	TCGA.HZ.7922	PAAD	C6	C1	**0.91**	0.00081	0.038	0.00056	0.00054	0.046	Yes
23	7284	TCGA.MS.A51U	BRCA	C6	C3	0.16	0.039	**0.57**	0.0063	0.005	0.22	No

*Being part of the classifier training set

Classifier performance (A1): balanced accuracy when using bootstrapping (1000 times) on training and validation sets



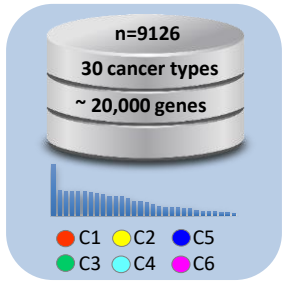


QUEEN'S UNIVERSITY BELFAST
SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

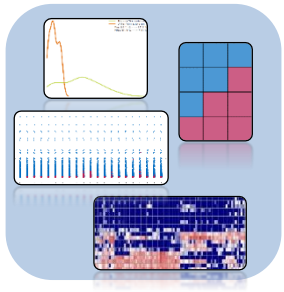
Immunogenomics subgrouping: training and prediction phases

Training phase

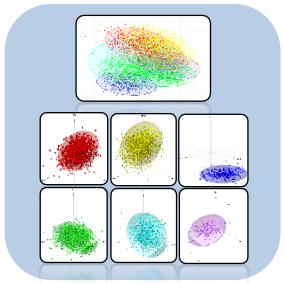
TCGA tumours
(RNA-seq with immune subgroup)



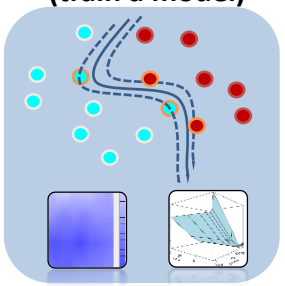
Pre-processing



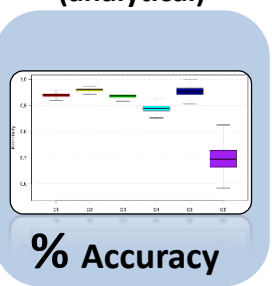
Feature preparation



Learning algorithm
(train a model)



Validation
(analytical)



Prediction phase

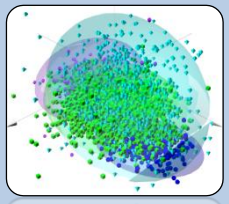
Unlabelled
gene expression data
(without immune subgroup)

- ~ 20,000 genes
- Multiple platforms
- Cross cancer (solid)
- Row or normalised
- FFPE or frozen tissues

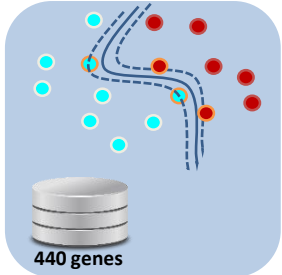
Pre-processing

- Harmonising
- Normalising
- Imputation

Feature preparation



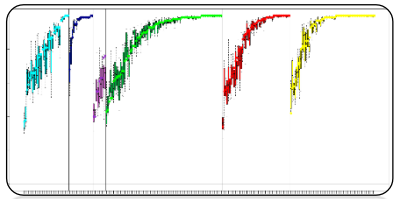
Trained model



Post-processing

- Probability
- Confidence interval
- Accuracy
- NC

Labelled data
(with subgroup)



http://smg.qub.ac.uk:3838/iga/

Any Questions?