Applied Machine Learning

Evaluation

Reihaneh Rabbany

McGill School of Computer Science

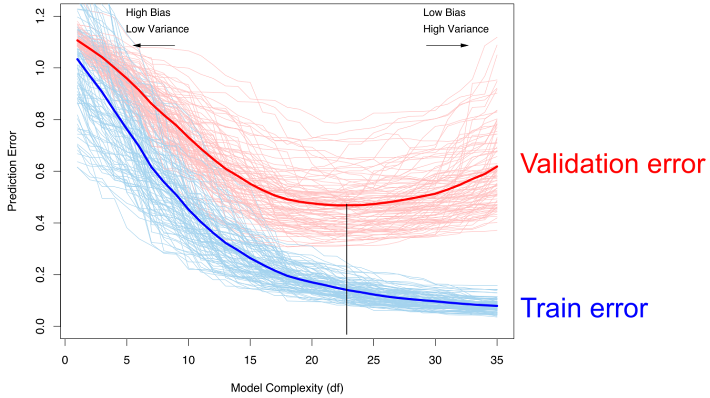
COMP 551 (winter 2020)

Evaluation and comparison

Given multiple models, how can we compare their performance?

We can report the loss function

e.g. least squares or cross entropy

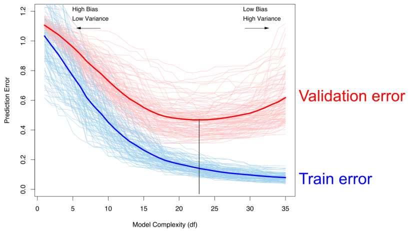


Evaluation and comparison

Given multiple models, how can we compare their performance?

We can report the loss function

e.g. least squares or cross entropy



What if each model is optimizing a different cost functions?

use standard evaluation measures/metrics

also more interpretable

Learning objectives

- different types of error
- common evaluation metrics
- cross validation

Performance metrics for classification

Not all errors are the same

In particular in classification, we have different types of mistakes

example:

false positive (type I) and false negative (type II)

patient does not have disease but received positive diagnostic (Type I error)
patient has disease but it was not detected (Type II error)

a message that is not spam is assigned to the spam folder (Type I error)

a message that is spam appears in the regular folder (Type II error)

Performance metrics for classification

classification results:

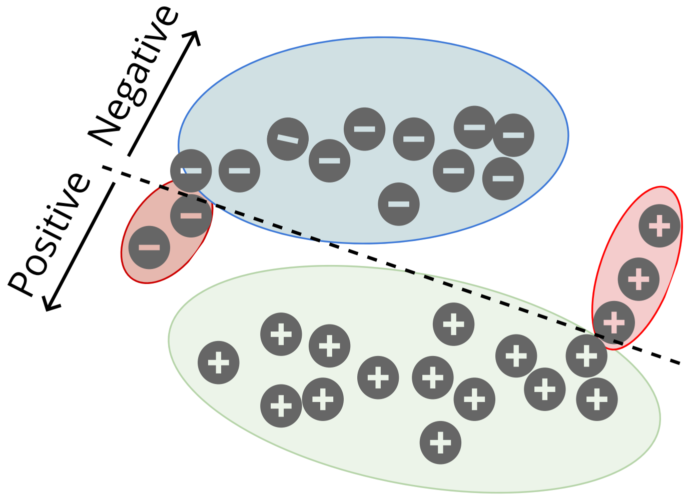
FP false positive (type I)

FN false negative (type II)

TP true positive

TN true negative

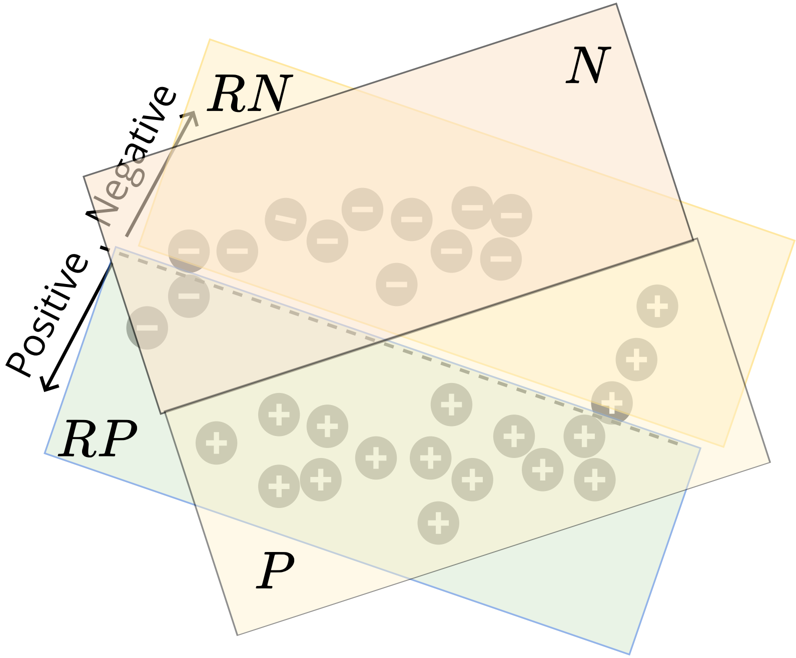
$$TN + TP + FN + FP = ?$$



Performance metrics for classification

confusion matrix

	Truth		Σ
Result	TP	FP	RP
	FN	TN	RN
Σ	P	N	



marginals of confusion matrix

$$RP = TP + FP$$
$$RN = TN + FN$$
$$P = TP + FN$$
$$N = TN + FP$$

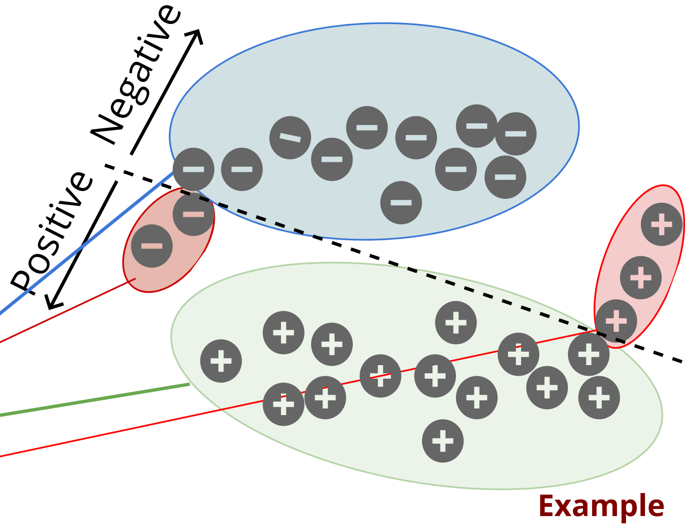
Performance metrics for classification

confusion matrix

	Truth		Σ
Result	TP	FP	RP
	FN	TN	RN
Σ	P	N	

example:

	Truth		Σ
Result	14	2	16
	3	11	14
Σ	17	13	



Performance metrics for classification

confusion matrix

	Truth		Σ
Result	TP	FP	RP
	FN	TN	RN
Σ	P	N	

marginals:

$$RP = TP + FP$$
$$RN = FN + TN$$
$$P = TP + FN$$
$$N = FP + TN$$

$$Accuracy = \frac{TP+TN}{P+N}$$

$$Error\ rate = \frac{FP+FN}{P+N}$$

$$Precision = \frac{TP}{RP}$$

$$Recall = \frac{TP}{P}$$

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

{Harmonic mean}

Performance metrics for classification

confusion matrix

	Truth		Σ
Result	TP	FP	RP
	FN	TN	RN
Σ	P	N	

$$Accuracy = \frac{TP+TN}{P+N}$$

$$Precision = \frac{TP}{RP}$$

$$Recall = \frac{TP}{P}$$

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

$$F_\beta score = (1+\beta^2)\frac{Precision \times Recall}{\beta^2 Precision + Recall}$$

recall is β times more important compared to precision

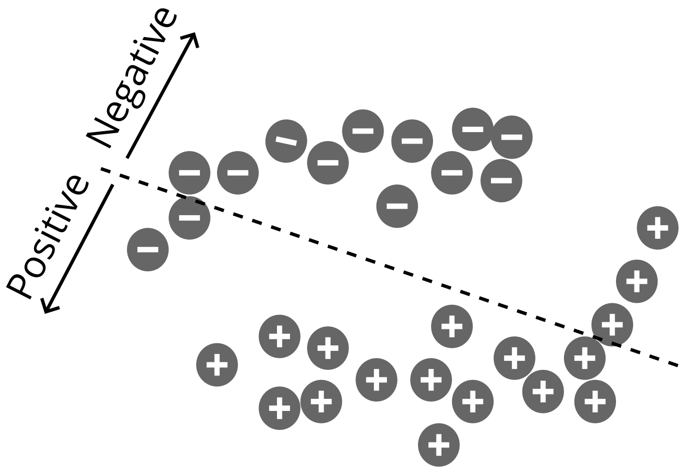
Performance metrics for classification

confusion matrix

	Truth		Σ
Result	TP	FP	RP
	FN	TN	RN
Σ	P	N	

example:

	Truth		Σ
Result	14	2	16
	3	11	14
Σ	17	13	



Example

$$Precision = \frac{TP}{RP} = \frac{14}{16}$$

$$Recall = \frac{TP}{P} = \frac{14}{17}$$

Performance metrics for classification

confusion matrix

	Truth		Σ
Result	TP	FP	RP
	FN	TN	RN
Σ	P	N	

$$Accuracy = \frac{TP+TN}{P+N}$$

$$Precision = \frac{TP}{RP}$$

$$Recall = \frac{TP}{P}$$ sensitivity

$$F_1 score = 2\frac{Precision \times Recall}{Precision + Recall}$$ {Harmonic mean}

Less common

$$Miss\ rate = \frac{FN}{P}$$

$$Fallout = \frac{FP}{N}$$ false positive rate

$$False\ discovery\ rate = \frac{FP}{RP}$$

$$Selectivity = \frac{TN}{N}$$ specificity

$$False\ omission\ rate = \frac{FN}{RN}$$

$$Negative\ predictive\ value = \frac{TN}{RN}$$

Performance metrics for multi class classification

confusion matrix

	Truth		Σ
Result	TP	FP	RP
	FN	TN	RN
Σ	P	N	

$$2 \times 2 \Rightarrow C \times C$$

report average metrics per class

e.g. average precision

$$M_{rc} = N\{\hat{y} = r, y = c\}$$

actual true classes

predicted results

	G	B	R	Y	Σ
G	0	0	0	0	0
B	12	6	0	0	18
R	0	0	11	0	11
Y	0	0	0	5	5
Σ	12	6	11	5	34

Measure the off diagonal density

more on this later in the course

Trade-off between precision and recall

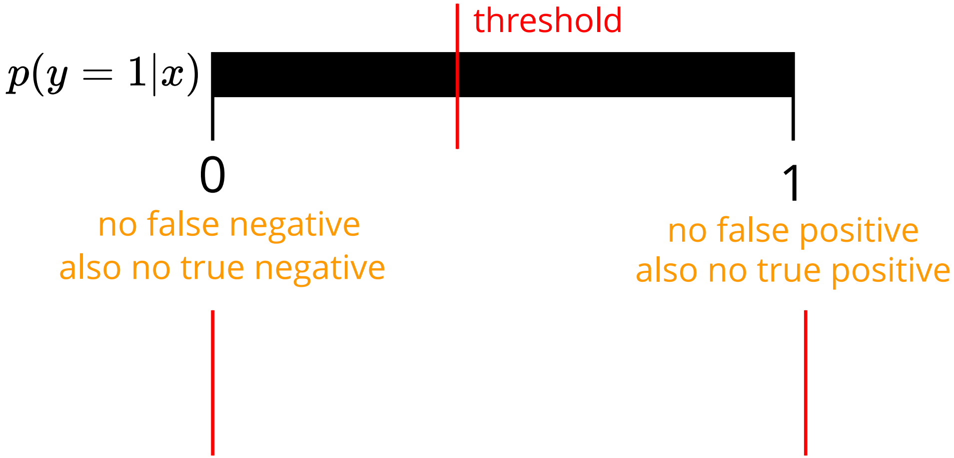
How many false positives do we tolerate?

How important are false negatives?

e.g. spam in inbox v.s. negative test for cancer test

We can often control the trade-off

e.g. by changing the threshold of $p(y=1|x)$ if we produce class score (probability)



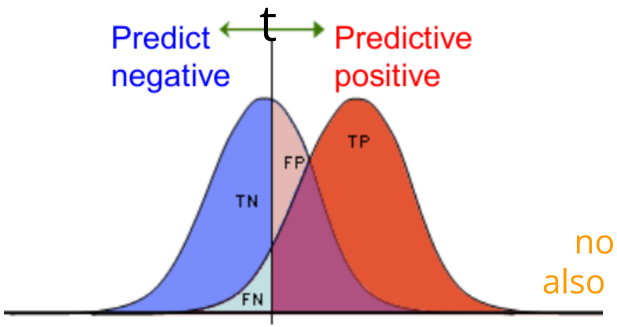
Threshold invariant: ROC & AUC

Receiver Operating Characteristic **ROC curve**

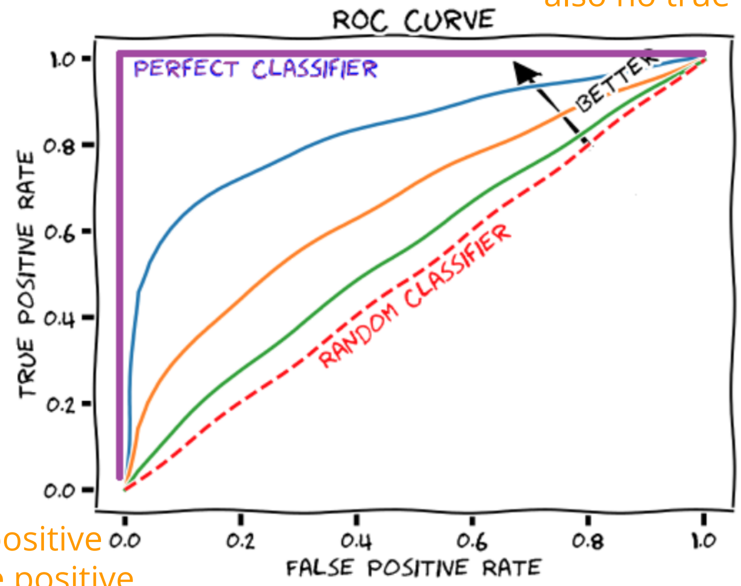
ROC as a function of threshold t

TPR(t) = TP(t)/P (**recall**, sensitivity at t)

FPR(t) = FP(t)/N (**fallout**, false alarm at t)



no false negative
also no true negative



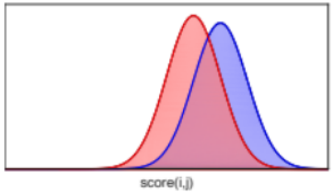
no false positive
also no true positive

Threshold invariant: ROC & AUC

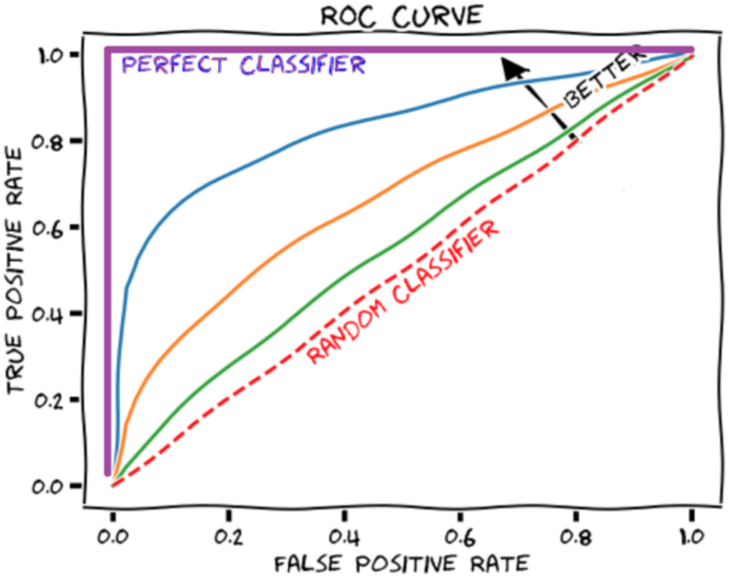
Receiver Operating Characteristic **ROC curve**

To compare classification algorithms compare their Area Under the Curve (**AUC**)

AUC = 0.99



AUC = 0.76



Higher AUC doesn't mean all performance measures are better

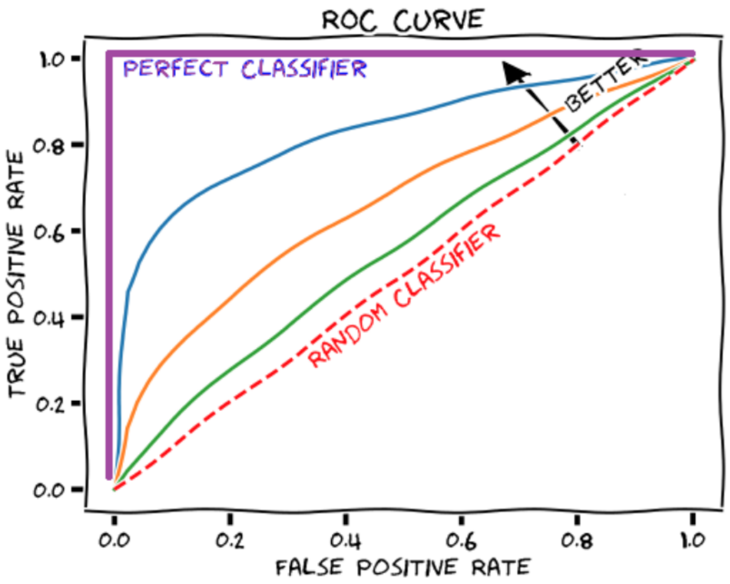
Threshold invariant: ROC & AUC

Also important when comparing ranking algorithms

e.g. search results

more on this later in the course

Intuition: **AUC** is equivalent to the probability of ranking a random positive example higher than a random negative example!



Model selection

how to pick the model with lowest expected loss / test error?

use a **validation set** (and a separate test set for final assessment)

use for model selection

use for final model assessment once the model is fixed and tuned

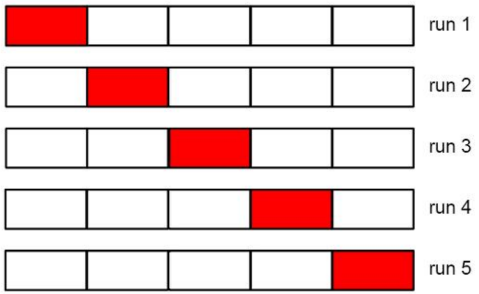
in the end we may have to use a validation set to find the right amount of regularization

Cross validation

getting a more reliable estimate of test error using validation set

K-fold cross validation(CV)

- randomly partition the data into K *folds*
- use K-1 for training, and 1 for validation
- report average/std of the validation error over all folds



increasing the folds gives better estimate of generalization error but takes more time and is k times more expensive to compute

leave-one-out CV: extreme case of k=N

Cross validation

Over-fitting in Model Selection

more severe on small dataset and when having too many hyper-parameters but present even with few hyperparameters

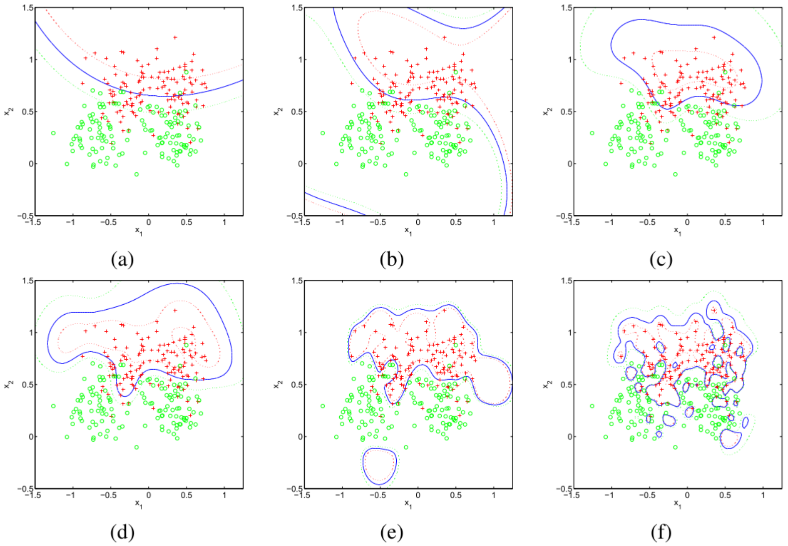


Figure 6: Kernel ridge regression models of the synthetic benchmark, using hyper-parameters selected according to the smoothed error rate over six random realisations of the validation set (shown in Figure 5). The variance of the model selection criterion can result in models ranging from under-fit, (a) and (b), through well-fitting, (c) and (d), to over-fit (e) and (f).

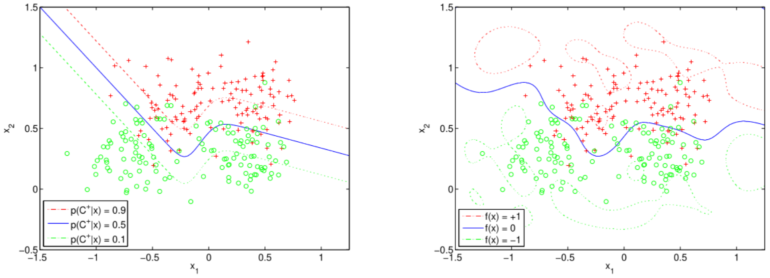
Credit: Cawley GC, Talbot NL. On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research. 2010;11(Jul):2079-107.

Cross validation

Cross validation is often used in:

> **evaluation** : Train & Test split
>
> **hyperparameter tuning** : Train & Validate split

use a single careful split for tuning or test, cross-validate on the other



(a) (b)

Figure 1: Realisation of the Synthetic benchmark data set, with Bayes optimal decision boundary (a) and kernel ridge regression classifier with an automatic relevance determination (ARD) kernel where the hyper-parameters are tuned so as to minimise the true test MSE (b).

Credit: Cawley GC, Talbot NL. On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research. 2010;11(Jul):2079-107.

nested Cross validation

Cross validation is often used in:

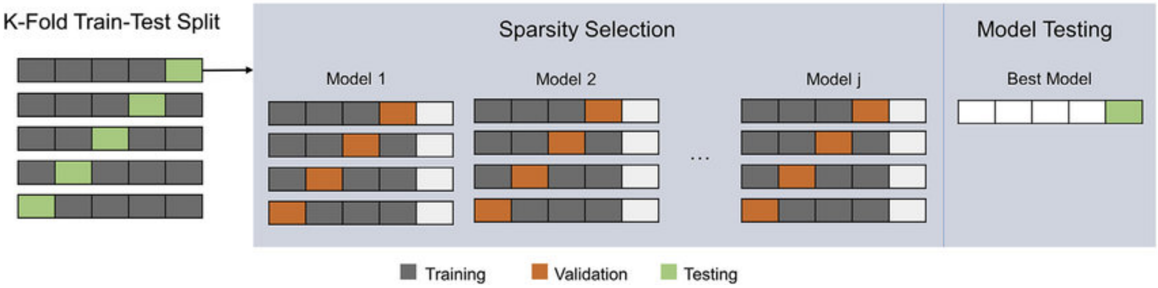
> **evaluation** : Train & Test split
>
> **hyperparameter tuning** : Train & Validate split

use a single careful split for tuning or test

nested cross validation

more rigorous but computationally intensive

Nested K-fold Cross-Validation with Model Selection



credit: figure from here

Accuracy not the only objective

examples:

Amazon's hiring algorithm decides not to invite women to interview.

Google's online ad algorithm decides to show high-income jobs to men much more often than to women.

Florida risk scores algorithm used in courts assign higher risk to black defendants

A health care algorithm offered less care to black patients

Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019 Oct 25;366(6464):447-53.

How can we factor these in the evaluation of models?

Summary

common measures of performance

ROC curves and AUC

cross-validation and model selection

fairness and bias also factors in evaluation