ML PERFORMANCE

JEROEN VEEN

CONTENTS

- Confusion matrix
- Evaluating classifiers
- Learning curves

THE BOY WHO CRIED WOLF

"Wolf" is a **positive class**.

"No wolf" is a **negative class**

An Aesop's Fable ~620 BCE

Source: Sam Taplin

CONFUSION MATRIX

ACTUAL

PREDICTED

(Type I error)

True Positive (TP)
Reality: A wolf threatened.
Shepherd said: "Wolf."
Outcome: Shepherd is a hero.

False Positive (FP)
Reality: No wolf threatened.
Shepherd said: "Wolf."
Outcome: Villagers are angry at shepherd for waking them up.

False Negative (FN)
Reality: A wolf threatened.
Shepherd said: "No wolf."
Outcome: The wolf ate all the sheep.

True Negative (TN)
Reality: No wolf threatened.
Shepherd said: "No wolf."
Outcome: Everyone is fine.

Type II error)
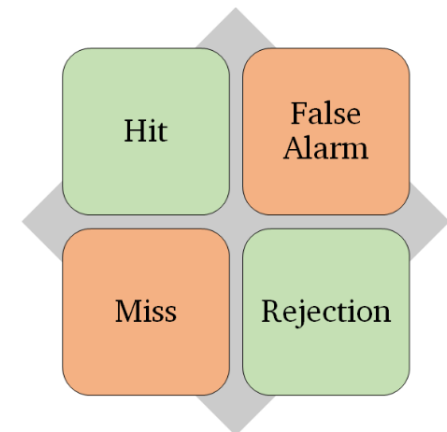
HAN_UNIVERSITY
OF APPLIED SCIENCES

ACCURACY

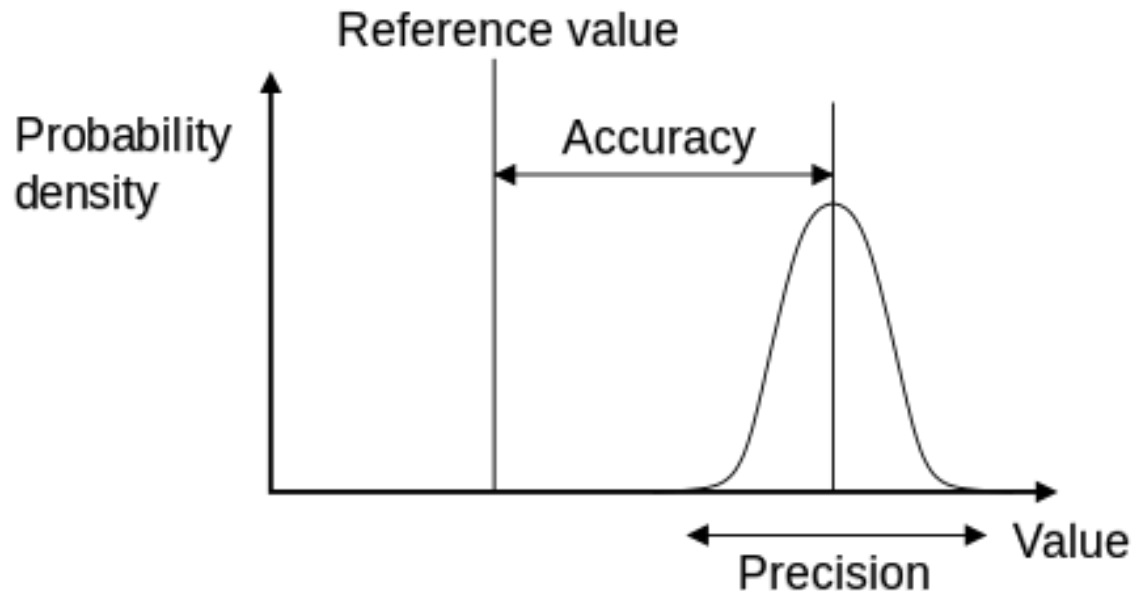- Fraction of predictions the model got right

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

- For binary classification

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Hit	False Alarm
Miss	Rejection

ACCURACY VS PRECISION

accuracy is closeness of the measurements to a specific value, while **precision** is the closeness of the measurements to each other.

HAN_UNIVERSITY
OF APPLIED SCIENCES
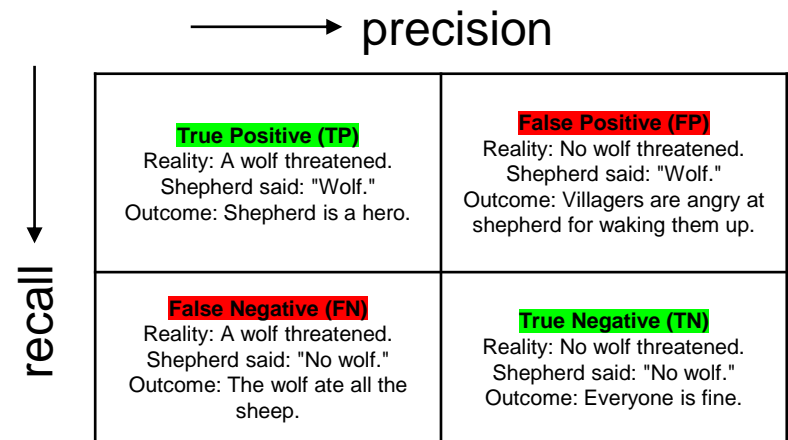
PRECISION AND RECALL

- Precision, fraction of correct positive predictions

  Σ True positive / Σ Predicted condition positive

$$\text{precision} = \frac{TP}{TP + FP}$$

- Recall, probability of detection
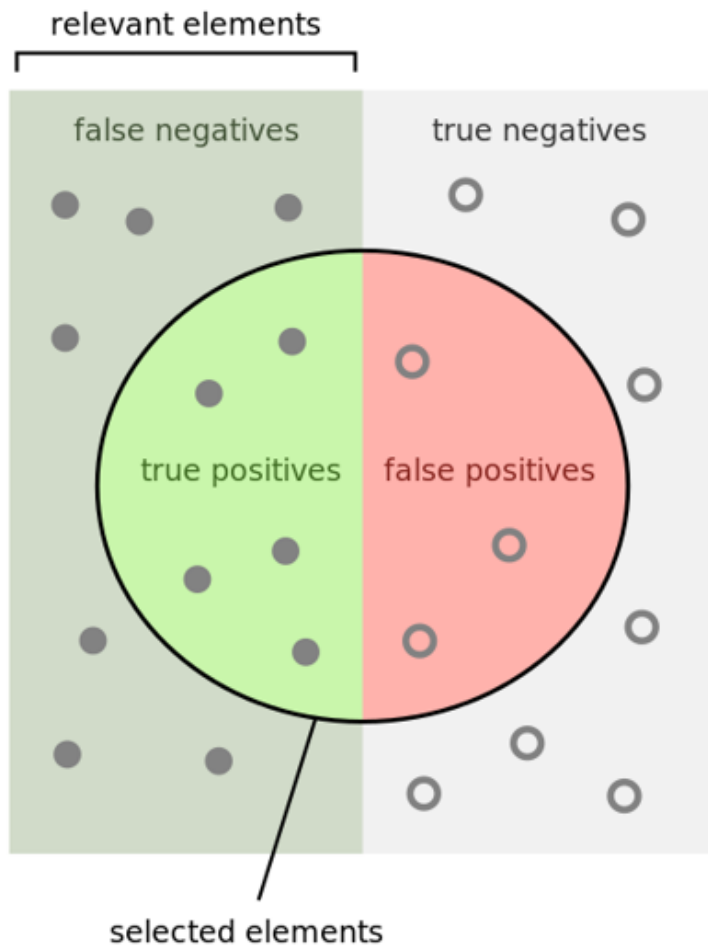
  Σ True positive / Σ Condition positive

$$\text{recall} = \frac{TP}{TP + FN}$$

precision →

recall ↓

True Positive (TP) Reality: A wolf threatened. Shepherd said: "Wolf." Outcome: Shepherd is a hero.	**False Positive (FP)** Reality: No wolf threatened. Shepherd said: "Wolf." Outcome: Villagers are angry at shepherd for waking them up.
False Negative (FN) Reality: A wolf threatened. Shepherd said: "No wolf." Outcome: The wolf ate all the sheep.	**True Negative (TN)** Reality: No wolf threatened. Shepherd said: "No wolf." Outcome: Everyone is fine.

PRECISION AND RECALL

relevant elements

| false negatives | true negatives |

true positives | false positives

selected elements

How many selected items are relevant?

$$Precision = \frac{\text{(green semicircle)}}{\text{(green+red circle)}}$$

How many relevant items are selected?

$$Recall = \frac{\text{(green semicircle)}}{\text{(green+dark green square)}}$$

Recall = sensitivity = true positive rate (TPR)

HAN_UNIVERSITY
OF APPLIED SCIENCES

LIMITATION OF USING SENSITIVITY AND SPECIFICITY
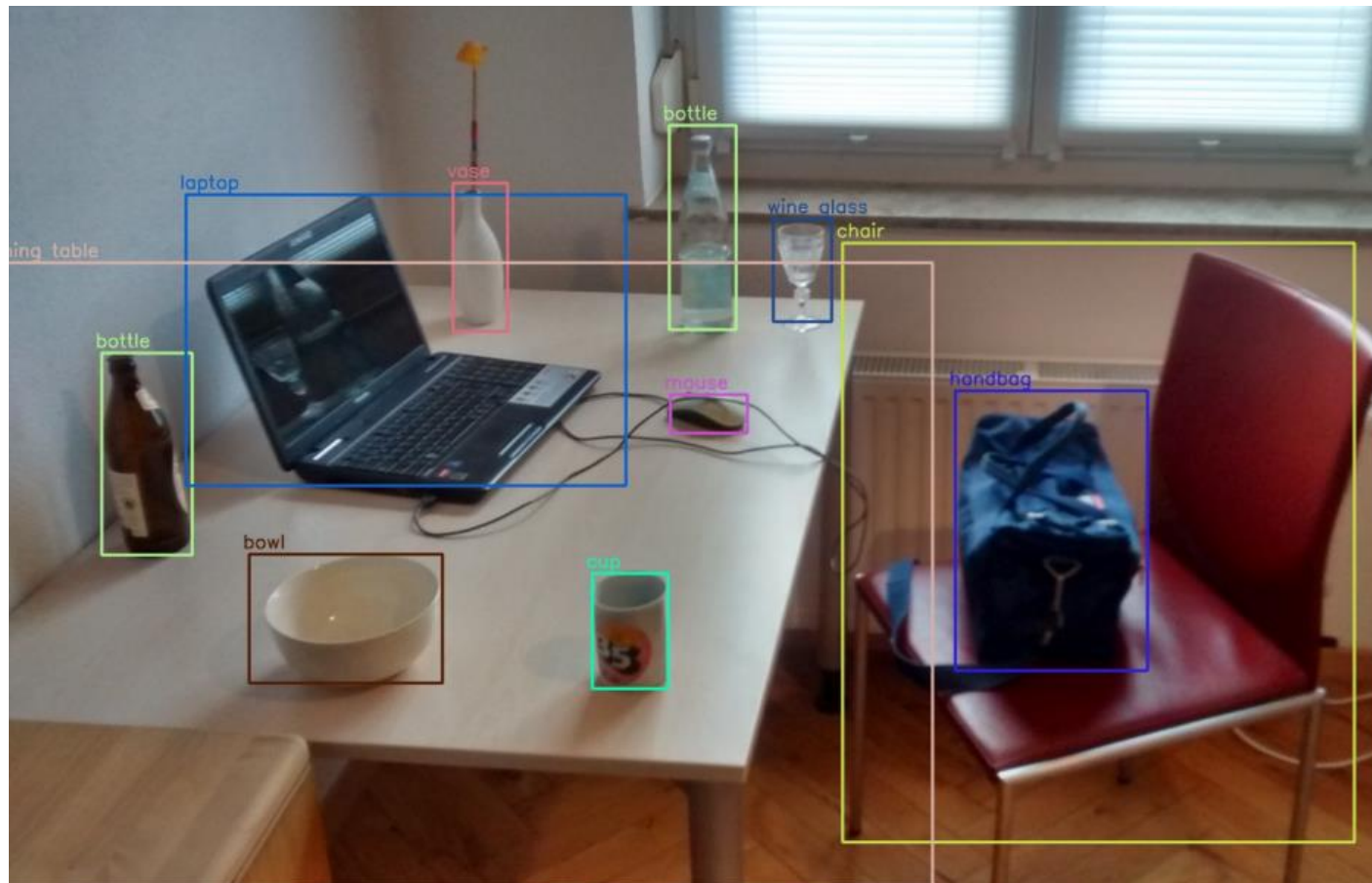
LET'S CHECK A VIDEO CLIP

- https://www.youtube.com/watch?v=TtIjAiSojFE

	precision	recall	f1-score
0.0	0.96	0.97	0.97
1.0	0.65	0.58	0.62

F1 SCORE

- To fully evaluate the effectiveness of a model, you must examine **both** precision and recall

- F1 score is the harmonic mean of precision and recall

$$F_1 = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{TP}{TP + \frac{FN + FP}{2}}$$

MANY METRICS

veren. Alle tests van de deelnemende partijen zijn eerst beoordeeld door het RIVM. „We sturen tien monsters, vijf met SARS-CoV-2 in verschillende concentraties, drie met andere coronavirussen, en twee waar niets in zit. De labs weten niet wat waar in zit. Als hun uitslagen kloppen, is de test goed", zegt Reusken.

De testprestaties zullen per lab wat variëren: de specificiteit (de kans dat je een niet-ziek iemand ook als niet-ziek detecteert) en de sensitiviteit (de kans dat je een ziek iemand als ziek detecteert). De E-gen test is gevoeliger dan de RdRP-gen test, maar wat minder specifiek: naar schatting 99 tot 99,5 procent. Tests op meerdere genen zijn iets specifieker. Door de variatie ligt die specificiteit landelijk minimaal op 99,5 procent, zegt Reusken.

4 | **Wat bepaalt hoeveel foute uitslagen een test geeft?**
„De specificiteit hangt af van de test zelf

Prevalence $= \dfrac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \dfrac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Positive predictive value (PPV), Precision $= \dfrac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \dfrac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
False omission rate (FOR) $= \dfrac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \dfrac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$

Positive likelihood ratio (LR+) $= \dfrac{\text{TPR}}{\text{FPR}}$

Negative likelihood ratio (LR−) $= \dfrac{\text{FNR}}{\text{TNR}}$

Diagnostic odds ratio (DOR) $= \dfrac{\text{LR+}}{\text{LR−}}$

F_1 score $= 2 \cdot \dfrac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

PRECISION/RECALL TRADE-OFF

Decision threshold
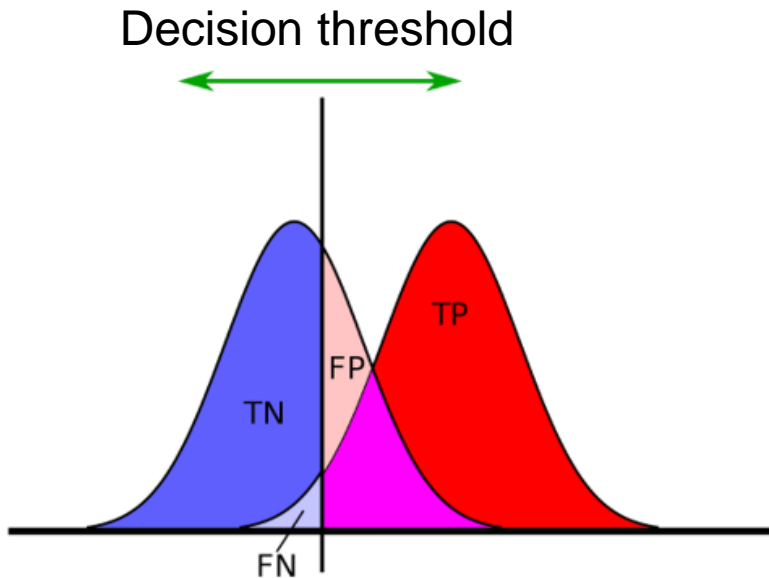
$$\text{precision} = \frac{TP}{TP + FP} \qquad \text{recall} = \frac{TP}{TP + FN}$$

TP	FP
FN	TN

Source: https://en.wikipedia.org/wiki/Tug_of_war#/media/File:Touwtrekken.jpg

HAN_UNIVERSITY OF APPLIED SCIENCES

ROC CURVE

- *probability of detection* vs *probability of false alarm* at different decision thresholds.
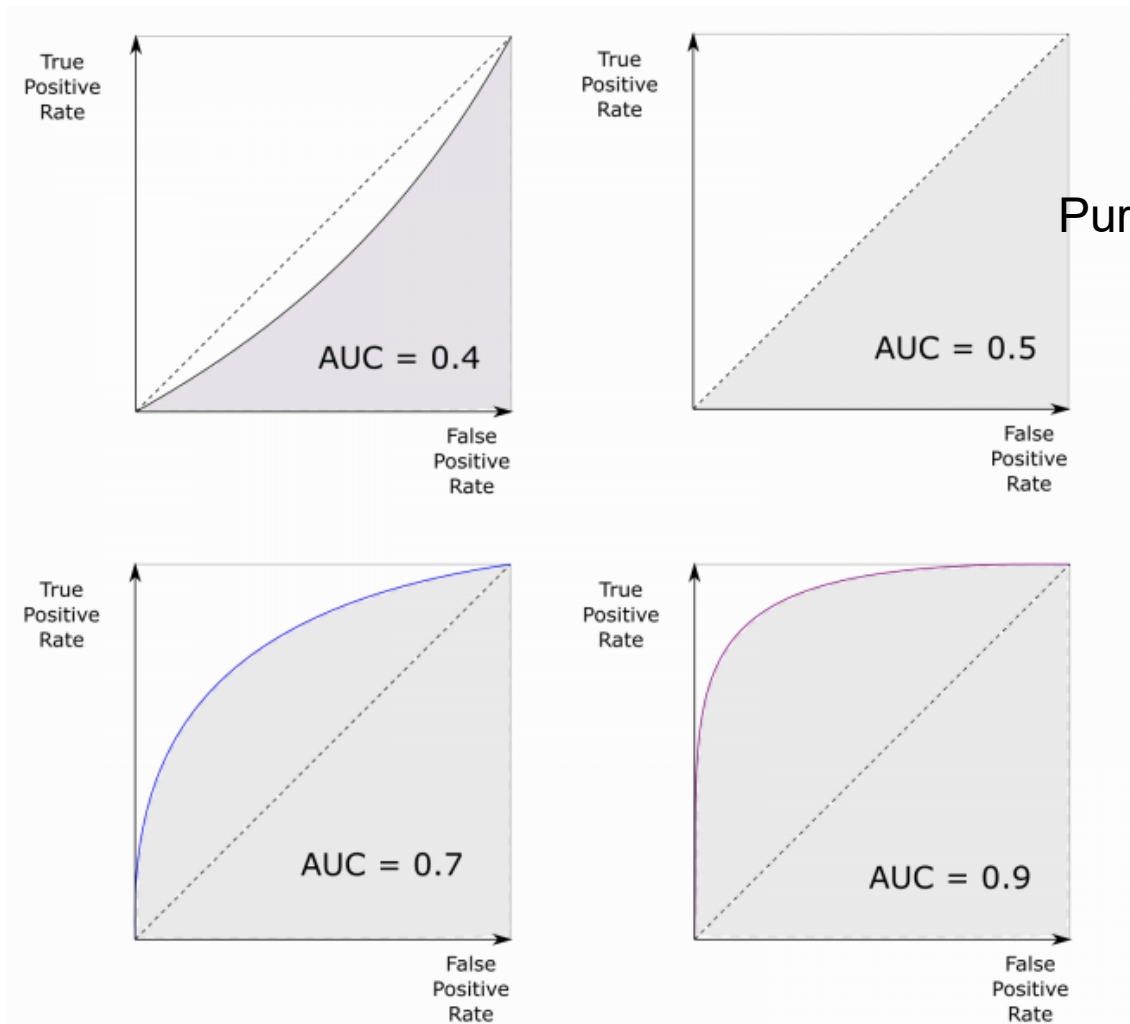

True positive rate
(precision)

False positive rate
(fall-out)
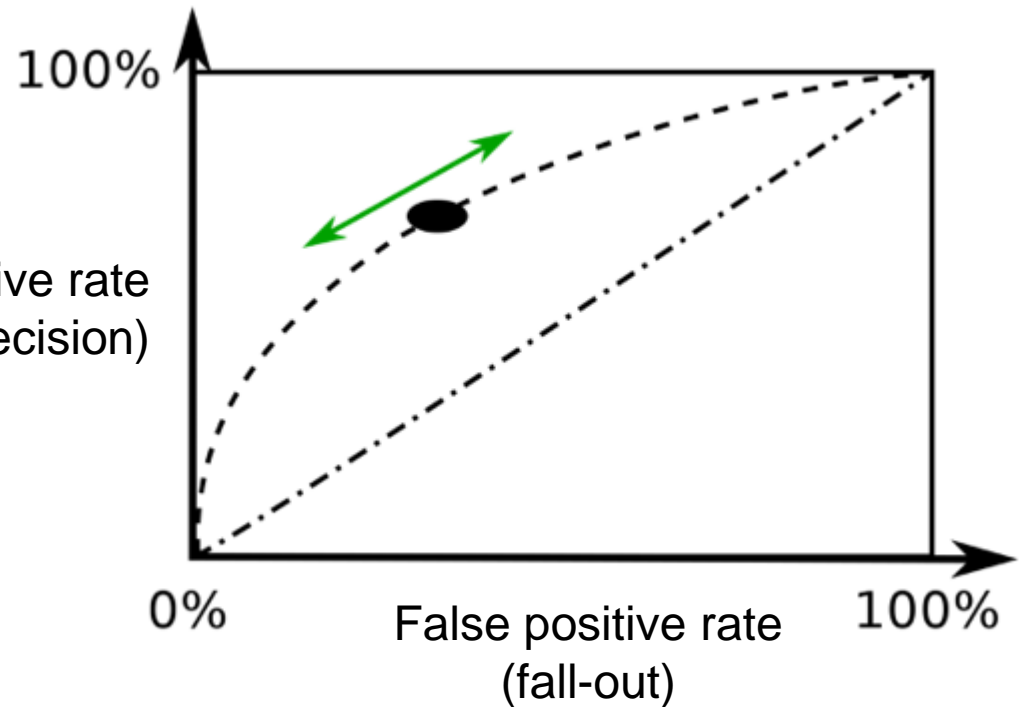
Source: https://en.wikipedia.org/wiki/Receiver_operating_characteristic#/media/File:ROC_curves.svg

COST OF CLASSIFICATION

- Sometimes false negatives don't hurt as much as false positives Think of a poisonous mushroom detector….

- Use the ROC curve (receiver operating characteristics) to help balance the cost of classification

ROC AREA UNDER THE CURVE (AUC)

Purely random…

MULTICLASS CONFUSION MATRIX

Source: https://miro.medium.com/max/1400/1*jtoE1zEJaG0JvGlX3jOTFQ.png

SPLITTING DATA

- Slice data into three subsets: Training, validation and test data

| ~60%	~20%	~20%
Training Set	Validation Set	Test Set

- Make sure that your subsets meet the following conditions:
  - Large enough to yield statistically meaningful results.
  - Representative of the data set as a whole.
    E.g. don't pick a test set with different characteristics than the training set.

HAN_UNIVERSITY
OF APPLIED SCIENCES

CROSS-VALIDATION

- Estimate of a model's generalization performance
- Break the data into folds

Source: https://www.kaggle.com/alexisbcook/cross-validation
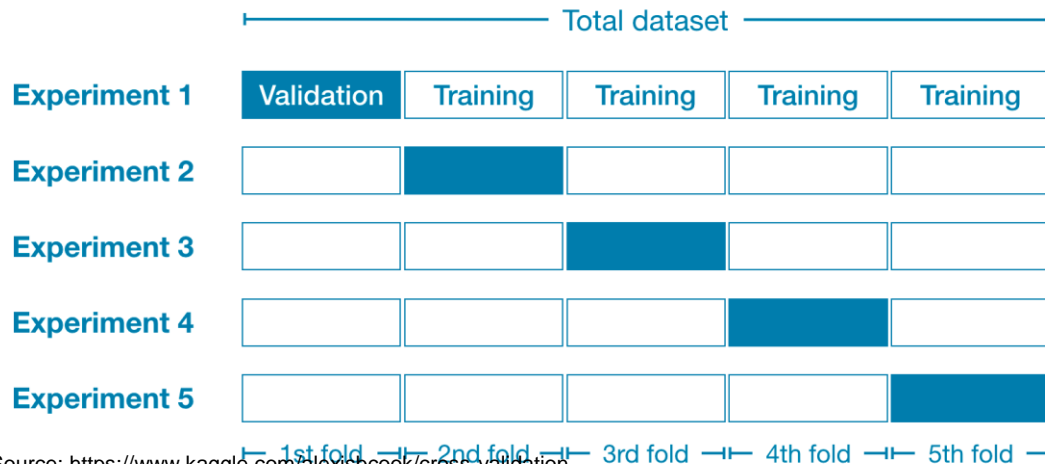
- For small datasets, where extra computational burden isn't a big deal, you should run cross-validation.
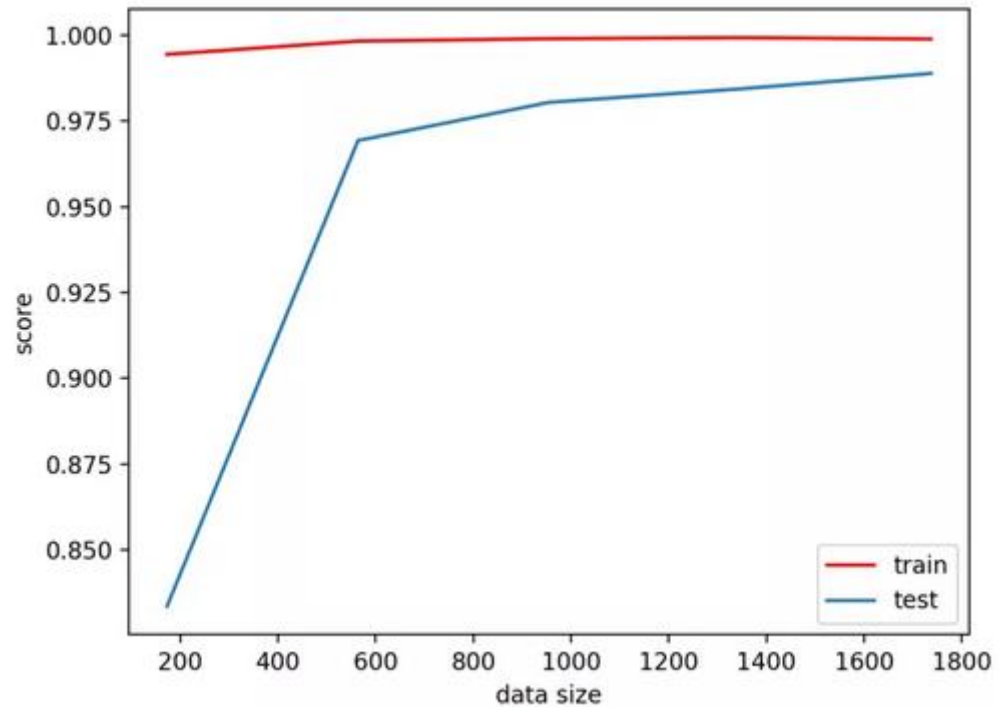
LEARNING CURVES


```python
from sklearn.model_selection import learning_c
from sklearn.svm import SVC
from sklearn.datasets import load_digits
from matplotlib import pyplot as plt
import numpy as np

X, y = load_digits(return_X_y=True)
estimator = SVC(gamma=0.001)

train_sizes, train_scores, test_scores, fit_times, _ = learning_curve(estimator, X, y, cv=30,return_times=True)

plt.plot(train_sizes,np.mean(train_scores,axis=1))
```
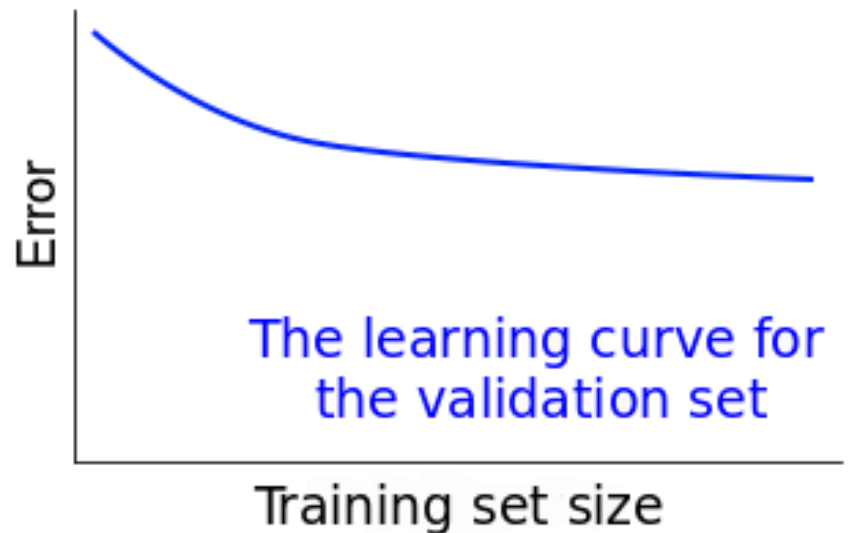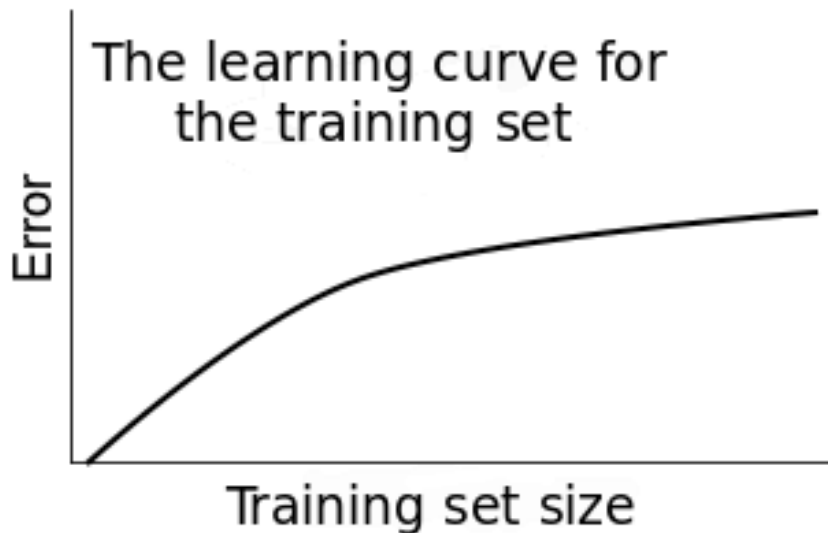
LEARNING CURVES
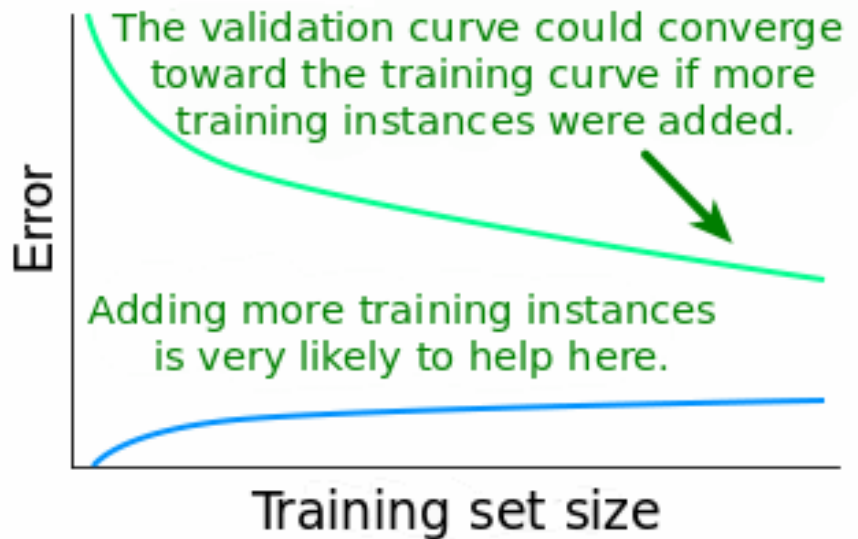
- Cost as a function of the training set size (or the training iteration)

- Examine evolution of train and validation learning curves

Source: https://www.dataquest.io/blog/learning-curves-machine-learning/

LEARNING CURVES

- Convergence of curves

- Diagnose generalization problems

Source: https://www.dataquest.io/blog/learning-curves-machine-learning/

BIAS PROBLEM

- High validation error indicates a prediction bias problem

- Underfitting usually gives high bias

Source: https://www.dataquest.io/blog/learning-curves-machine-learning/
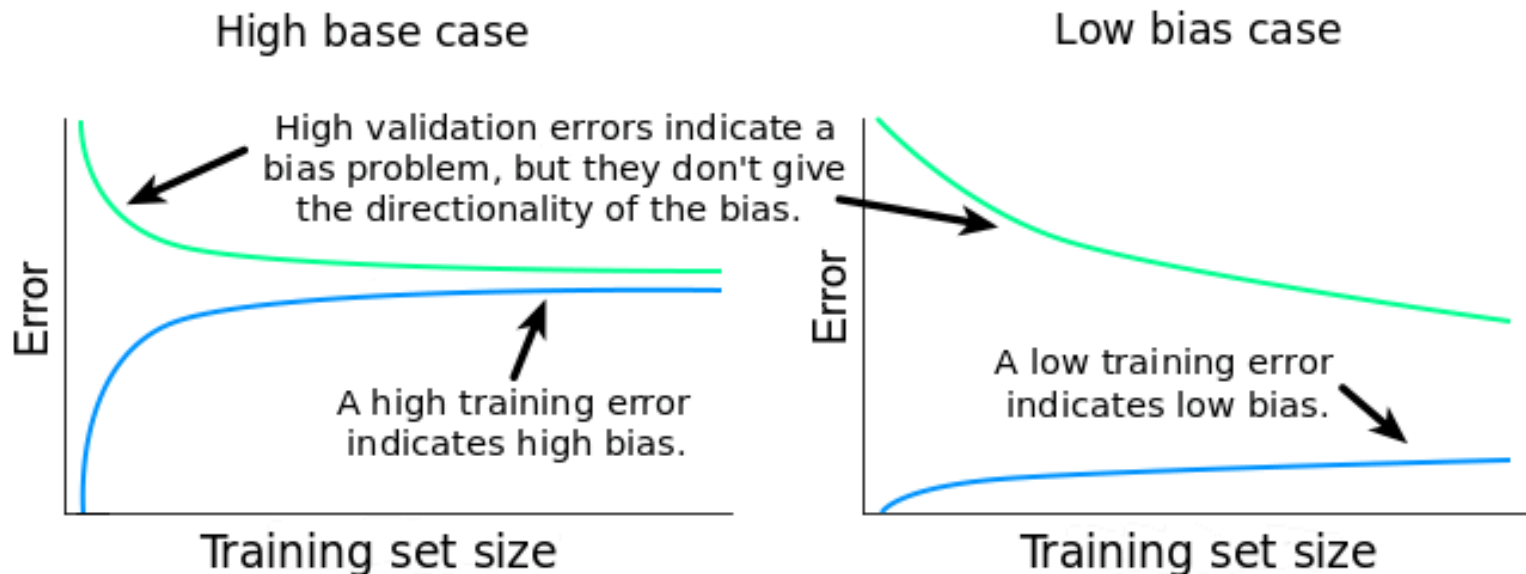
VARIANCE PROBLEM

- Low gap indicates low prediction variance

- Overfitting usually gives high variance

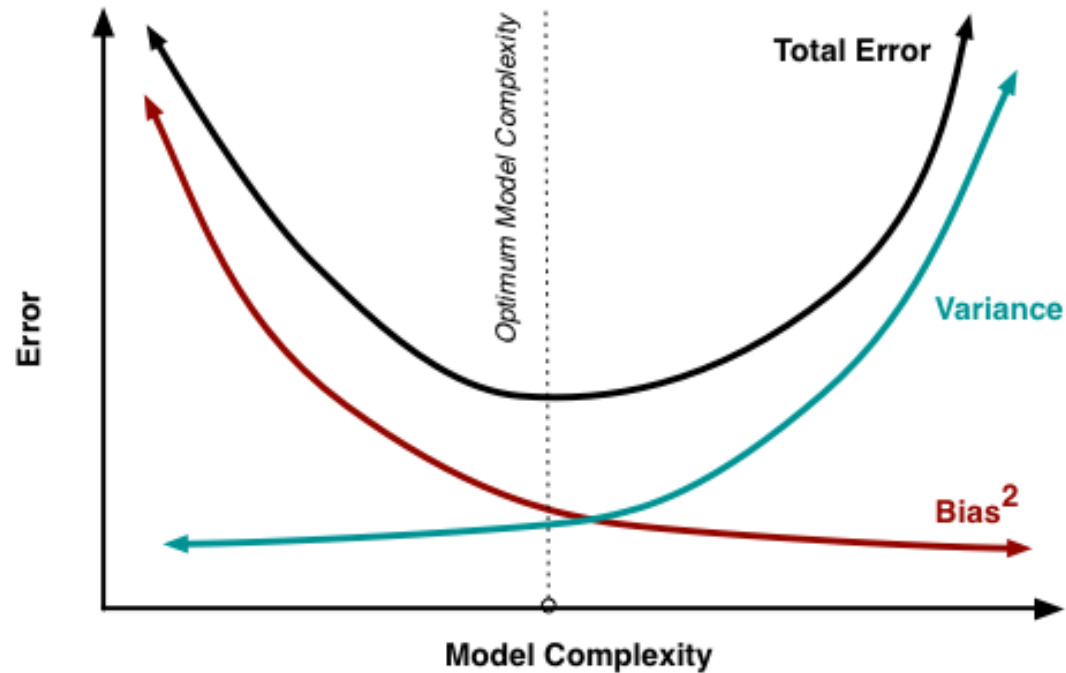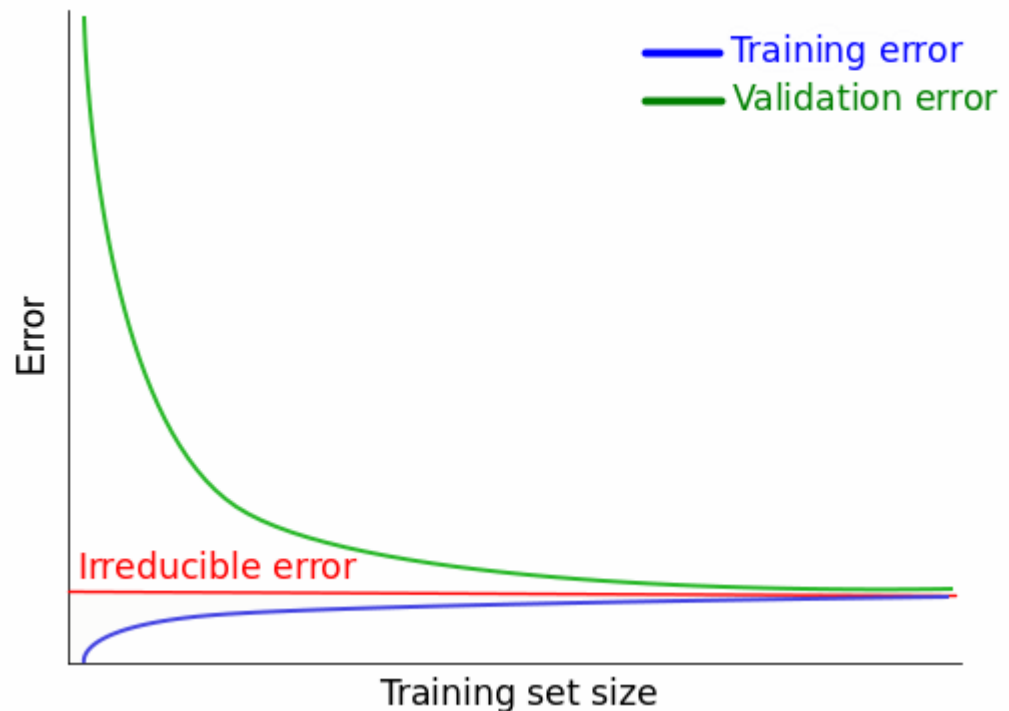Source: https://www.dataquest.io/blog/learning-curves-machine-learning/
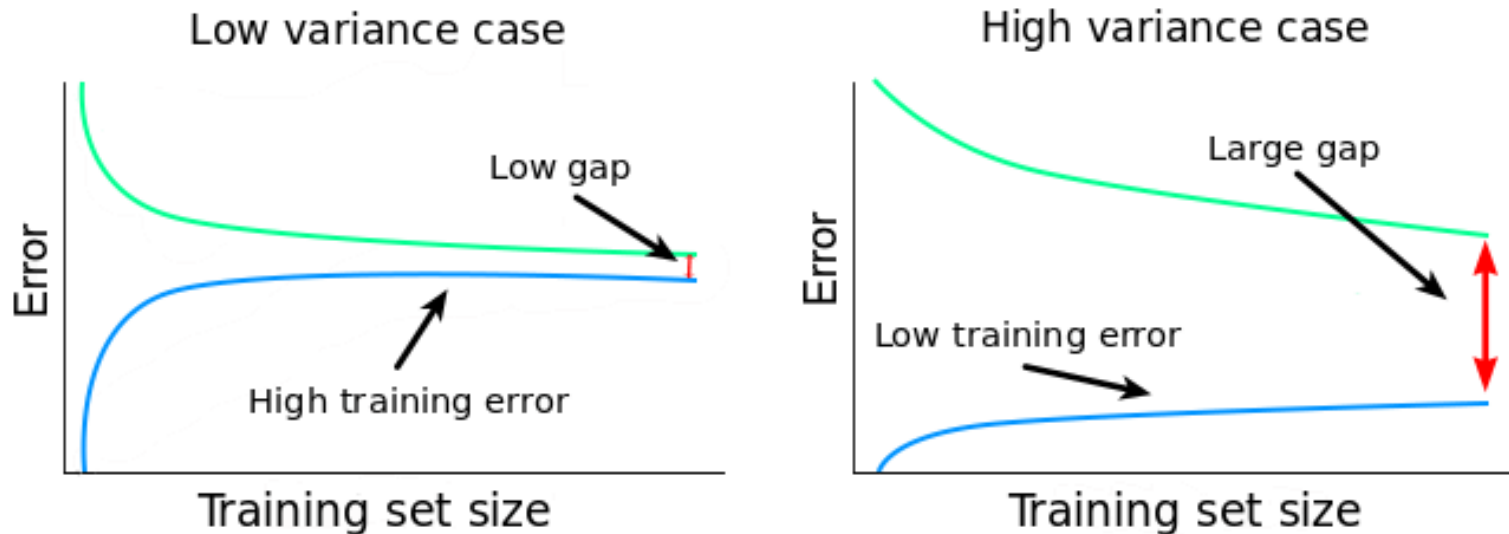
PREDICTION BIAS–VARIANCE TRADEOFF

- Central problem in supervised learning

IRREDUCIBLE ERROR

- Noise

- Cannot be predicted

- Loss cannot be reduced

HAN_ UNIVERSITY
OF APPLIED SCIENCES

MORE ON PREDICTION BIAS

- Average of predictions ≈ average of labels in test set

- A significant difference shows there is bias

- Possible causes:
  - Underfitting, e.g. incomplete feature set, overly strong regularization
  - Biased training samples
  - (Noisy data set)

QUESTION

- We know that on average, 1% of all emails are spam.

- My spam filter predicts that 20% of my incoming mail is spam.

What can we say about my spam filter?

COMPUTING CROSS-VALIDATED METRICS

• Predefined scoring parameters

Scoring	Function	Comment
Classification		
'accuracy'	metrics.accuracy_score	
'balanced_accuracy'	metrics.balanced_accuracy_score	
'average_precision'	metrics.average_precision_score	
'neg_brier_score'	metrics.brier_score_loss	
'f1'	metrics.f1_score	for binary targets
'f1_micro'	metrics.f1_score	micro-averaged
'f1_macro'	metrics.f1_score	macro-averaged
'f1_weighted'	metrics.f1_score	weighted average
'f1_samples'	metrics.f1_score	by multilabel sample
'neg_log_loss'	metrics.log_loss	requires predict_proba support
'precision' etc.	metrics.precision_score	suffixes apply as with 'f1'
'recall' etc.	metrics.recall_score	suffixes apply as with 'f1'
'jaccard' etc.	metrics.jaccard_score	suffixes apply as with 'f1'
'roc_auc'	metrics.roc_auc_score	
'roc_auc_ovr'	metrics.roc_auc_score	

See: https://scikit-learn.org/stable/modules/model_evaluation.html

HAN_UNIVERSITY
OF APPLIED SCIENCES