Foundations of Machine Learning
CentraleSupélec — Fall 2017

4. Model evaluation & selection

Chloé-Agathe Azencott
Centre for Computational Biology, Mines ParisTech
chloe-agathe.azencott@mines-paristech.fr

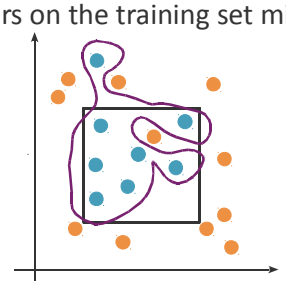
Generalization

A good and useful approximation

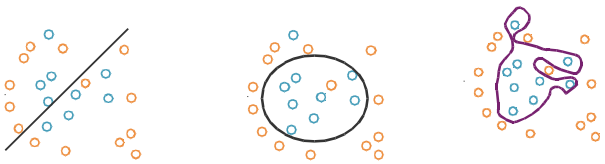
- It's easy to build a model that performs well on the training data
- But how well will it perform on **new data?**
- "Predictions are hard, especially about the future" — Niels Bohr.
  - Learn models that **generalize** well
  - Evaluate whether models generalize well.

Noise in the data

- Imprecision in **recording the features**
- **Errors in labeling** the data points (**teacher noise**)
- **Missing features** (**hidden** or **latent**)
- Making no errors on the training set might not be possible.

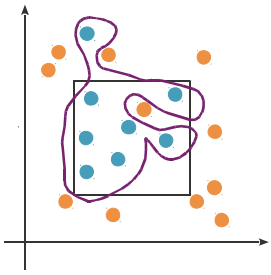


Models of increasing complexity

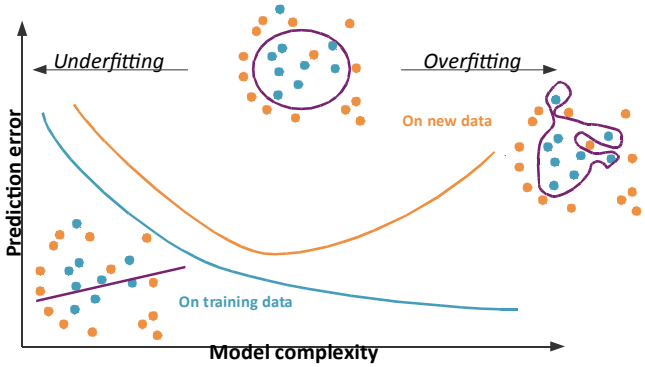


Noise and model complexity

- **Use simple models!**
  - Easier to **use**
    lower computational complexity
  - Easier to **train**
    lower space complexity
  - Easier to **explain**
    more interpretable
  - **Generalize better**
    Occam's razor: simpler explanations are more plausible.



Generalization error vs. model complexity



Bias-variance tradeoff

- **Bias:** difference between the expected value of the estimator and the true value being estimated.

$$\text{Bias}(f(\boldsymbol{x})) = \mathbb{E}[f(\boldsymbol{x}) - y]$$

  - A simpler model has a higher bias.
  - **High bias can cause underfitting.**
- **Variance:** deviation from the expected value of the estimates.

$$\text{Var}(f(\boldsymbol{x})) = \mathbb{E}[(f(\boldsymbol{x}) - \mathbb{E}(f(\boldsymbol{x}))^2]$$

  - A more complex model has a higher variance.
  - **High variance can cause overfitting.**

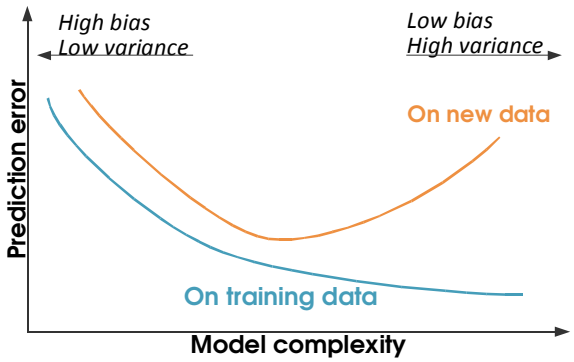
Bias-variance decomposition

- $\text{Bias}(f(\boldsymbol{x})) = \mathbb{E}[f(\boldsymbol{x}) - y]$
- $\text{Var}(f(\boldsymbol{x})) = \mathbb{E}[(f(\boldsymbol{x}) - \mathbb{E}(f(\boldsymbol{x}))^2]$
- **Mean squared error:**

$$\begin{aligned} \text{MSE}(f(\boldsymbol{x})) &= \mathbb{E}[(f(\boldsymbol{x}) - y)^2] \\ &= \text{Var}(f(\boldsymbol{x})) + \text{Bias}^2(f(\boldsymbol{x})) \end{aligned}$$

- Proof 

Generalization error vs. model complexity



Model selection & generalization

- **Well-posed problems:**
  - a solution exists;
  - it is unique;
  - the solution changes continuously with the initial conditions

Hadamard, on the mathematical modelisation of physical phenomena.

- Learning is an **ill-posed problem**:

  data helps carve out the hypothesis space

  but data is not sufficient to find a unique solution.

- Need for **inductive bias**

  assumptions about H

  model selection: choose the "right" inductive bias?

How do we decide a model is good?

Learning objectives

After this lecture you should be able to

design experiments to select and evaluate supervised machine learning models.

Concepts:

- training and testing sets;
- cross-validation;
- bootstrap;
- measures of performance for classifiers and regressors;
- measures of model complexity.

Supervised learning setting

- **Training set:** $\mathcal{D} = \{\boldsymbol{x}^i, y^i\}_{i=1,\ldots,n}$
- **Classification:** $y^i \in$
- **Regression:** $y^i \in$
- Goal: Find $f \in \mathcal{F}$ such that $f(\boldsymbol{x}^i) \approx y^i$
- **Empirical error** of f on the training set, given a **loss**:
  - E.g. (classification)
  - E.g. (regression)

Validation sets

- Choose the model that performs best on a **validation set separate from the training set**.

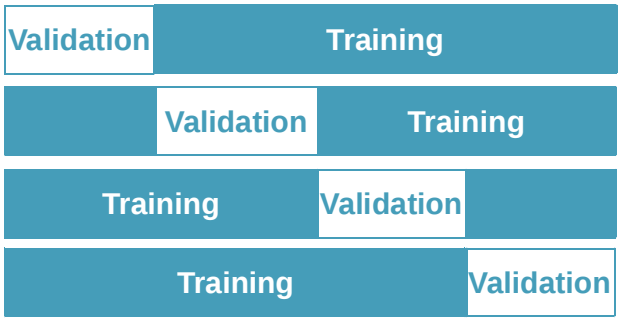
- Model **selection:** pick the best model.
- Model **assessment:** estimate its prediction error on new data.

- **How much data** should go in each of the training, validation and test sets?
- How do we know we have **enough data** to evaluate the prediction and generalization errors?
- **Sample re-use**
  - cross-validation
  - bootstrap
- **Analytical tools**
  - Mallow's Cp, AIC, BIC
  - MDL.

Sample re-use

Cross-validation

- Cut the training set in k separate **folds.**
- For each fold, train on the (k-1) remaining folds.



Cross-validated performance

- Cross-validation estimate of the prediction error

$$CV(f) = \frac{1}{n}\sum_{i=1}^{n} L(y^i, f_{k(i)}(\boldsymbol{x}^i))$$

Computed with the k(i)-th part of the data removed. k(i) = fold in which i is.

- Estimates the **expected prediction error**

$$\mathrm{Err} = \mathbb{E}[L(Y, f(X))]$$

Y, X: (independent) test sample

Issues with cross-validation

- **Training set size** becomes (K-1)n/K
- **Leave-one-out cross-validation:** K = n
  - approximately **unbiased estimator** of the expected prediction error
  - potential **high variance** (the training sets are very similar to each other)
  - **computation** can become burdensome (n repeats)
- In practice: set **K = 5 or K = 10.**

Bootstrap

- **Randomly draw datasets with replacement** from the training data
- **Repeat B times** (typically, B=100) ⇒ B models
- **Leave-one-out bootstrap error:**
  - For each training point i, predict with the $b_i < B$ models that did not have i in their training set
  - Average prediction errors
- Each training set contains

Evaluating model performance

Classification model evaluation

- **Confusion matrix**

		True class	
		-1	+1
Predicted class	-1	True Negatives	False Negatives
	+1	False Positives	True Positives

- False positives (false alarms) are also called **type I errors**
- False negatives (misses) are also called **type II errors**

- **Sensitivity** = **Recall** = True positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

positives

- **Specificity** = True negative rate (TNR)

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}}$$

- **Precision** = Positive predictive value (PPV)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

predicted positives

- **False discovery rate** (FDR)

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TP}}$$

- **Accuracy**

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN}}$$

- **F1-score** = harmonic mean of precision and sensitivity.

$$\mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

Example: Pap smear

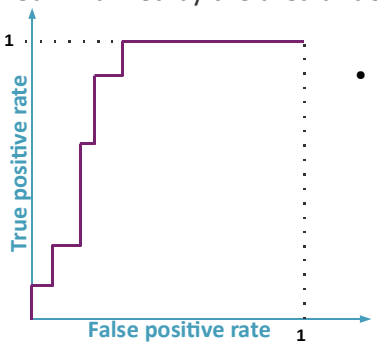
- 4,000 apparently healthy women of age 40+
- Tested for cervical cancer through pap smear and histology (gold standard)

	Cancer	No cancer	Total
Positive test	190	210	400
Negative test	10	3590	3600
Total	200	3800	4000

- **What are the sensitivity, specificity, and PPV of the test?**

ROC curves

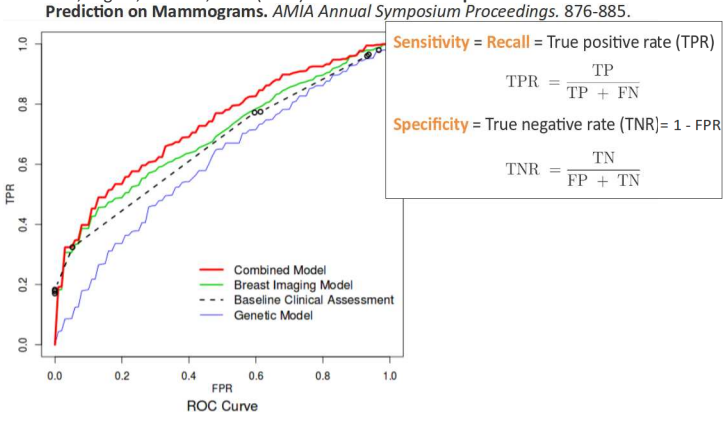
- ROC = Receiver-Operator Characteristic.
- Summarized by the area under the curve (AUROC).



- Plot TPR vs FPR for all possible thresholds.

Predicting breast cancer risk based on mammography images, SNPs, or both.

Liu J, Page D, Nassif H, et al. (2013). **Genetic Variants Improve Breast Cancer Risk Prediction on Mammograms.** *AMIA Annual Symposium Proceedings.* 876-885.



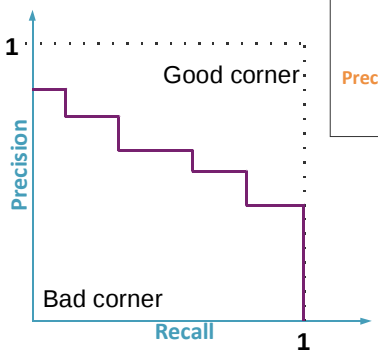
Sensitivity = Recall = True positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Specificity = True negative rate (TNR) = 1 - FPR

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}}$$

Precision-Recall curves



Sensitivity = **Recall** = True positive rate (TPR)

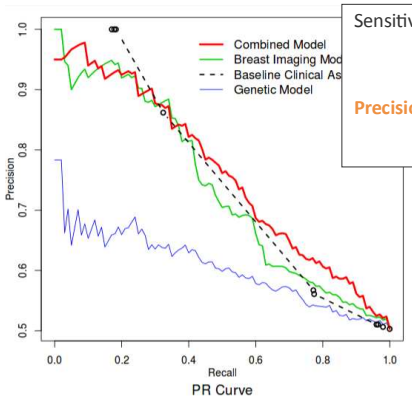
$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Precision = Positive predictive value (PPV)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Predicting breast cancer risk based on mammography images, SNPs, or both.

Liu J, Page D, Nassif H, et al. (2013). **Genetic Variants Improve Breast Cancer Risk Prediction on Mammograms.** *AMIA Annual Symposium Proceedings.* 876-885.



Sensitivity = **Recall** = True positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Precision = Positive predictive value (PPV)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Regression model evaluation

- Counting the number of errors is not reasonable

Regression model evaluation

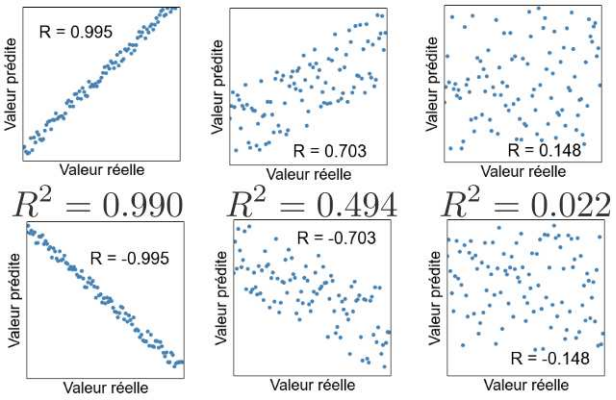
- **Residual sum of squares** $\mathrm{RSS} = \sum_{i=1}^{n} (y^i - f(\boldsymbol{x}^i))^2$
- **Root-mean squared error**

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y^i - f(\boldsymbol{x}^i))^2}$$

- **Relative squared error** $\mathrm{RSE} = \frac{\sum_{i=1}^{n} (y^i - f(\boldsymbol{x}^i))^2}{\sum_{i=1}^{n} (y^i - \bar{y})^2}$
- **Coefficient of determination**

$$R^2 = 1 - \mathrm{RSE} = \frac{\sum_{i=1}^{n} (y^i - \bar{y})(f(\boldsymbol{x}^i) - \overline{f(\boldsymbol{x})})}{\sqrt{\sum_{i=1}^{n} (y^i - \bar{y})^2} \sqrt{\sum_{i=1}^{n} (f(\boldsymbol{x}^i) - \overline{f(\boldsymbol{x})})}}$$

Correlation between true and predicted values



Analytical tools and model complexity

Optimism terms

- Correct the empirical error with an **optimism term**
- Theoretical estimate of the **discrepancy between training and test error**

Augmented error = empirical error + optimism term

- For **linear models**, optimism terms proportional to:
  - **Mallow's Cp:** $\frac{\hat{\sigma}^2}{n} d$
    - $\hat{\sigma}^2$: Variance of the residuals on the train set
    - $\frac{1}{n}$: Squared standard error of the mean of the residuals
    - d: # parameters = # non-zero coefficients
  - **Akaike Information Criterion (AIC):** d
  - **Bayesian Information Criterion (BIC):** $d \ln(n)$

Minimum description length (MDL)

- **Shortest code to transmit a random variable z:**

  $-\log_2 P(\boldsymbol{z})$ [Shannon's source coding theorem]

- Assume
  - Parametric model f_θ
  - receiver knows inputs X, model family f.
- To transmit outputs y, need

$$\underbrace{-\log_2 P(\boldsymbol{y}|\theta, f, X)}_{\text{average code length to transmit the difference between model prediction and true outputs.}} \underbrace{-\log_2 P(\theta)}_{\text{average code length to transmit θ.}}$$

- Choose the model with smallest Kolmogorov complexity (=MDL)

Summary: model selection techniques

- **Empirical:**

  Estimate quality of generalization with
  - **cross-validation**
  - **bootstrap**
- **Theoretical:**
  - Estimate the difference between train error and generalization error with an optimism term

    E.g. Mallow's Cp, Akaike's / Bayesian Information Criteria
  - **Minimum description length (MDL)**

    Choose simplest model (according to Kolmogorov complexity)

References

- *A Course in Machine Learning.* `http://ciml.info/dl/v0_99/ciml-v0_99-all.pdf`
  - **Noise:** Chap 2.3
  - **Overfitting:** Chap 2.4
  - **Bias-variance tradeoff:** Chap 5.9
  - **Train and test sets:** Chap 2.5
  - **Cross-validation:** Chap 5.6
  - **Performance measures:** Chap 5.5
- *The Elements of Statistical Learning.* `http://web.stanford.edu/~hastie/ElemStatLearn/`
  - **Overfitting:** Chap 7.1
  - **Bias-variance tradeoff:** Chap 2.9, 7.2–7.3
  - **Cross-validation:** Chap 7.10
  - **Bootstrap:** Chap 7.11
  - **Mallow's Cp, AIC, BIC:** Chap 7.7
  - **MDL:** Chap 7.8
- **Entropy encoding:** `http://lesswrong.com/lw/o1/entropy_and_short_codes/`