Lecture 15. Cross-Validation

COMP90051 Statistical Machine Learning

Lecturer: Jean Honorio

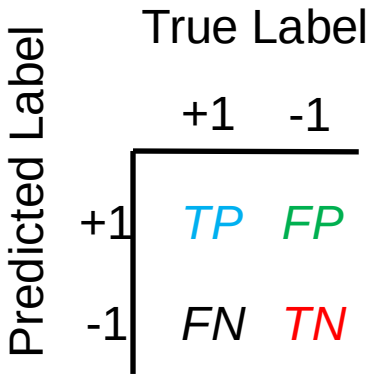
THE UNIVERSITY OF MELBOURNE

Performance metrics

- Given a dataset $D = \{x_1, y_1, \ldots, x_n, y_n\}$ of n samples
- Assume that for a data point x_i we predict $g(x_i)$
- Some metrics in regression:
  - Mean squared error: $MSE(g) = \frac{1}{n}\sum_{i=1}^{n}(g(x_i) - y_i)^2$
  - Root mean squared error: $RMSE(g) = \sqrt{MSE(g)}$
  - Mean absolute error: $\frac{1}{n}\sum_{i=1}^{n}|g(x_i) - y_i|$

Performance metrics

- True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN)



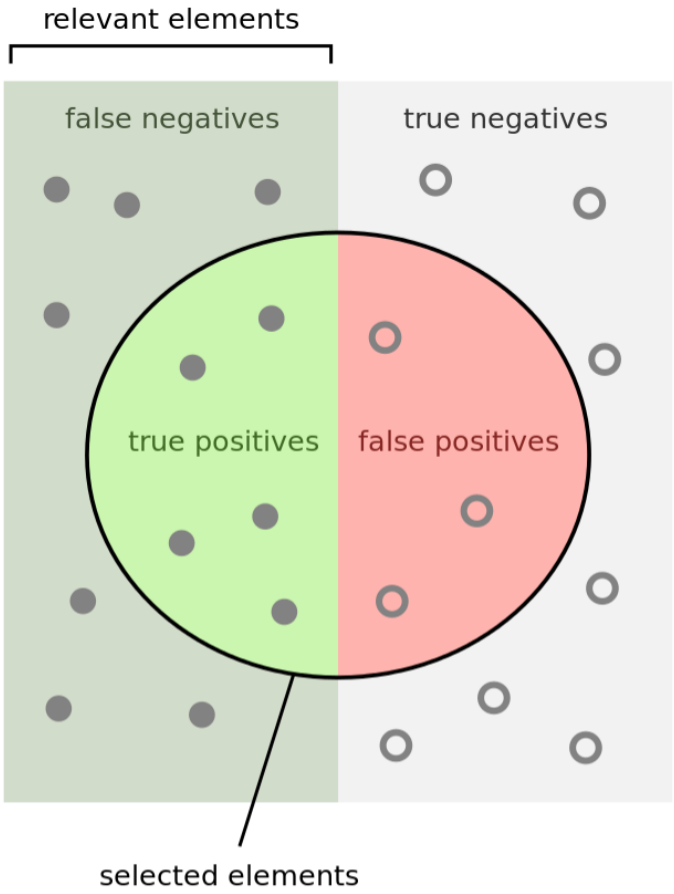
- Some metrics classification:
  - Accuracy $(TP + TN)/(TP + FP + FN + TN)$
  - Error $(FP + FN)/(TP + FP + FN + TN)$
  - Recall / Sensitivity $TP/(TP + FN)$
  - Precision $TP/(TP + FP)$
  - Specificity $TN/(TN + FP)$
  - F1-score $2\ Precision \times Recall/(Precision + Recall)$

Performance metrics

- Precision and recall are typically in an inverse relationship:
  - The classifier has high Precision, but low Recall
  - The classifier has high Recall, but low Precision
- Similar for sensitivity and specificity
- Use jointly:
  - (Precision,Recall)
  - (Sensitivity,Specificity)

Precision and recall

- Idea comes from information retrieval



How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{selected elements}}$$

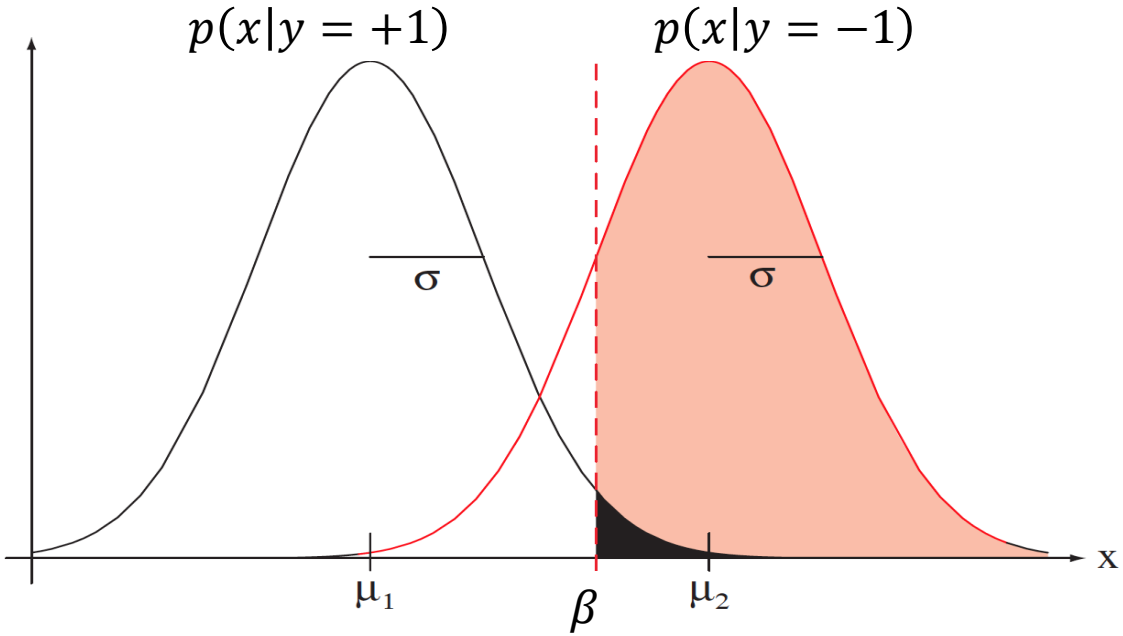
How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{relevant elements}}$$

Wikipedia

Sensitivity and specificity

- Idea comes from signal detection theory
  - Assume Gaussian distributions $p(x|y=+1) = N(\mu_1, \sigma^2)$
    $p(x|y=-1) = N(\mu_2, \sigma^2)$



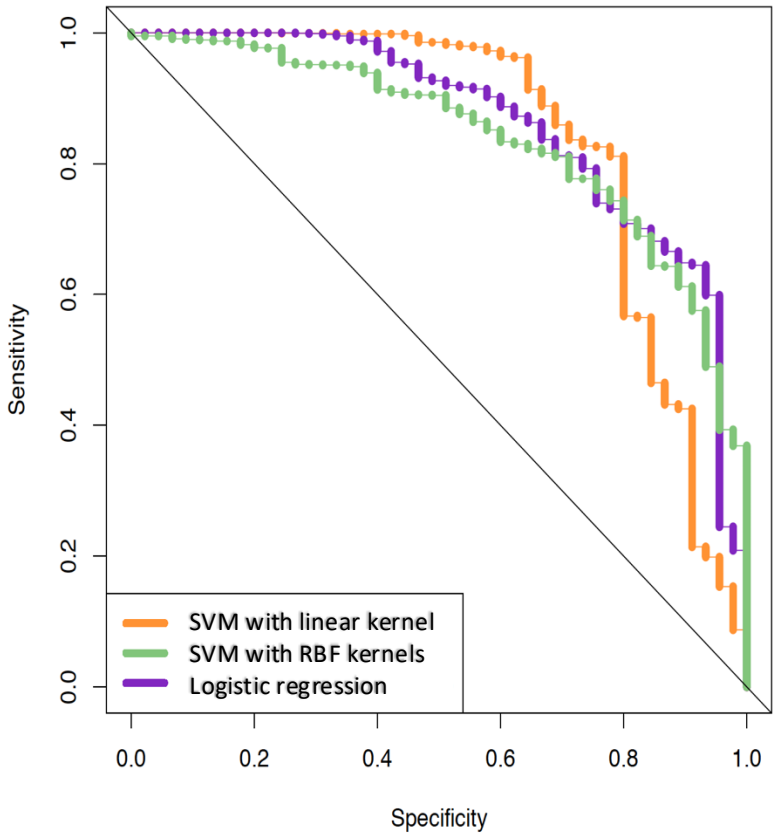
Classifier:

$$g(x) = \begin{cases} +1, \text{if } x < \beta \\ -1, \text{if } x \geq \beta \end{cases}$$

  - By sliding the offset β we get different (TP, FP, TN, FN) and thus, different sensitivity and specificity

Receiver Operating Characteristic (ROC)

- By varying the offset (or a hyperparameter) for a classifier (e.g., SVMs, logistic regression) we can get different:
  - Sensitivity
  - Specificity
- Summarized with an Area Under the Curve (AUC)
  - Random: 0.5
  - Perfect classifier: 1



The Elements of Statistical Learning, Hastie et al.

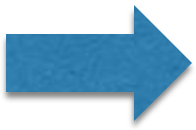
Chance level error

- Consider a classification problem:
  - 2 classes, 50% of samples each class
- Consider three classifiers:
  - A classifier that always predicts class +1 gets 50% error
  - A classifier that always predicting class -1 gets 50% error
  - A random classifier that would flip a coin on every prediction gets 50% error (approximately)
- Can a classifier get more than 50% error?
  - Yes, for test error (e.g., we could get less than 50/100 marks in an exam with true/false questions)

Other loss functions

- Consider classification with health data
  - Let +1 be “diseased patient” and -1 be “healthy patient”
  - Predicting a healthy patient as diseased (e.g., unnecessary medical procedures) has different consequences than predicting a diseased patient as healthy (e.g., missing early treatment)

Predicted Label \ True Label	+1	-1
+1	0	1
-1	1	0



Predicted Label \ True Label	+1	-1
+1	0	1
-1	10	0

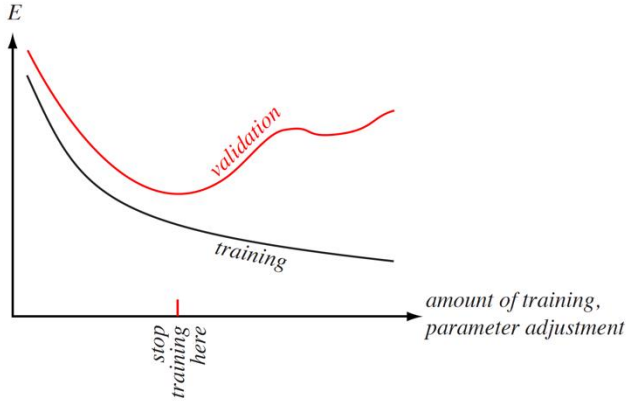
0/1 loss:

$$\frac{1}{n}\sum_{i=1}^{n} 1[g(x_i) \neq y_i]$$

$$\frac{1}{n}\sum_{i=1}^{n} Cost(g(x_i), y_i)$$

Using "unseen" data

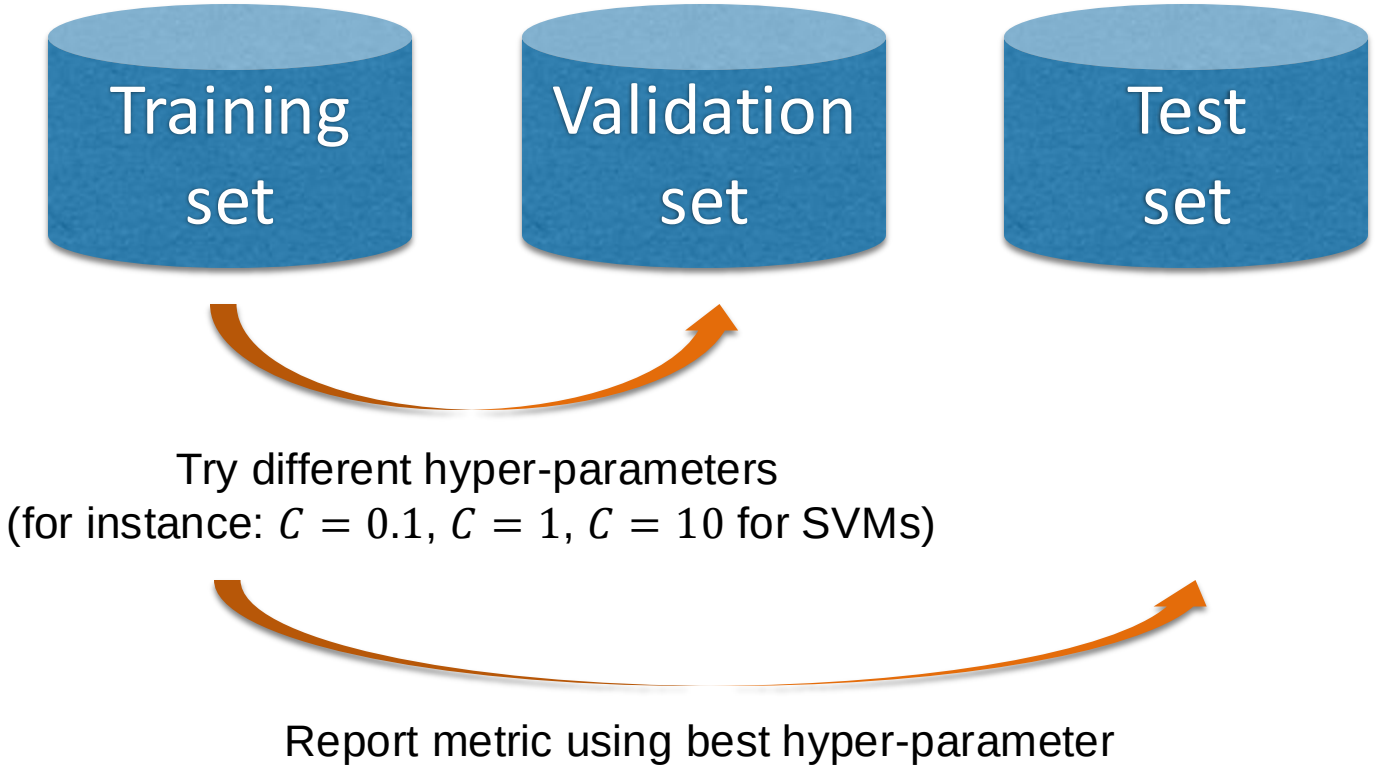
- Overfitting:
  - More complex methods fit better the training data (e.g., linear kernel versus cubic kernel)
  - Some hyper-parameter values might also fit better the training data
  - Usually poor performance in unseen data



- To prevent overfitting, how to "see" unseen data?
  - Simulate it !

Training, validation, testing

- Split datasets in three parts



- We can and should repeat this several times
  - (discussion on variability coming next)

k-Fold Cross Validation

- Split training data D into k disjoint sets $S_1, \ldots, S_k$
  - Either randomly, or in a fixed fashion
  - If D has n samples, then each fold has approximately n/k samples
  - Popular choices: $k = 5$, $k = 10$, $k = n$ (leave-one-out)
- For $i = 1 \ldots k$:
  - train with sets $S_1, \ldots, S_{i-1}, S_{i+1}, \ldots, S_k$
  - test on set S_i
  - let M_i be the test metric (e.g., accuracy, MSE)
- Mean $\hat{\mu} = \sum_{i=1}^{k} M_i$ variance $\hat{\sigma}^2 = \sum_{i=1}^{k} (M_i - \hat{\mu})^2$

0.632 Bootstrapping

- Let $B > 0$, and n be the number of training samples in D
- For $i = 1 \ldots B$:
  - Pick n samples from D with replacement, call it S_i (S_i might contain the same sample more than once)
  - train with set S_i
  - test on the remaining samples ($D - S_i$)
  - let M_i be the test metric (e.g., accuracy, MSE)
- Mean $\hat{\mu} = \sum_{i=1}^{B} M_i$ variance $\hat{\sigma}^2 = \sum_{i=1}^{B} (M_i - \hat{\mu})^2$

0.632 Bootstrapping

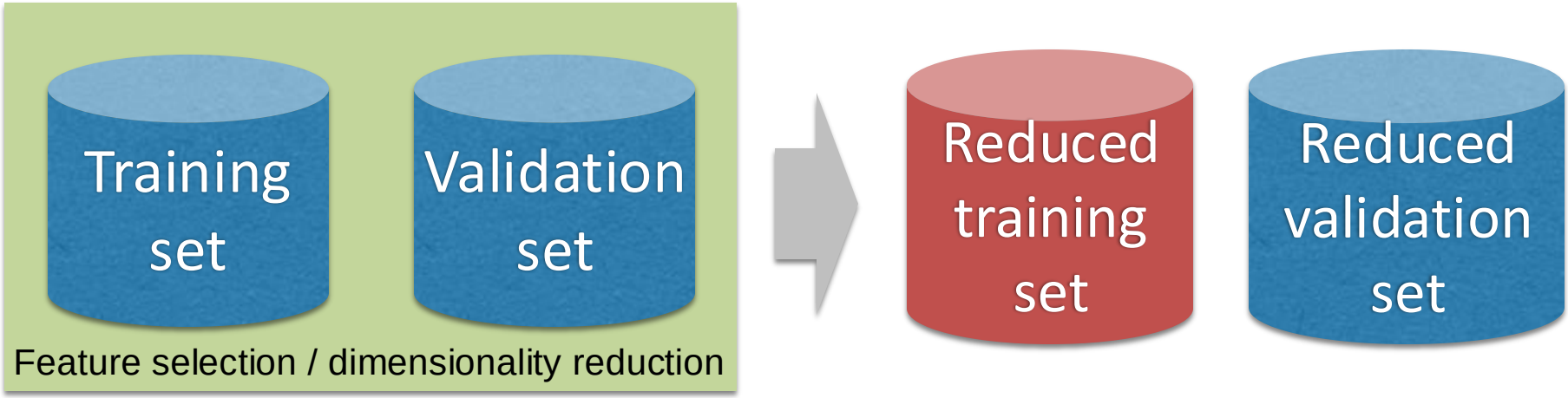
- Why 0.632?
- Recall that:
  - We pick n items with replacement from out of n items
  - We choose uniformly at random
- The probability of:
  - not picking one particular item in 1 draw is $1 - 1/n$
  - not picking one particular item in n draws is $(1 - 1/n)^n$
  - picking one particular item in n draws is $1 - (1 - 1/n)^n$
- Finally: $\lim_{n \to \infty} 1 - (1 - 1/n)^n = 1 - 1/e \approx 0.632$

Nested cross-validation

- Useful to choose the best model, for instance, for hyperparameter tuning
- Inner cross-validation (e.g., 0.632 bootstrapping) to try different hyperparameters
- Outer cross-validation (e.g., k-folds) to report metric using the best hyperparameter

Feature selection and cross-validation

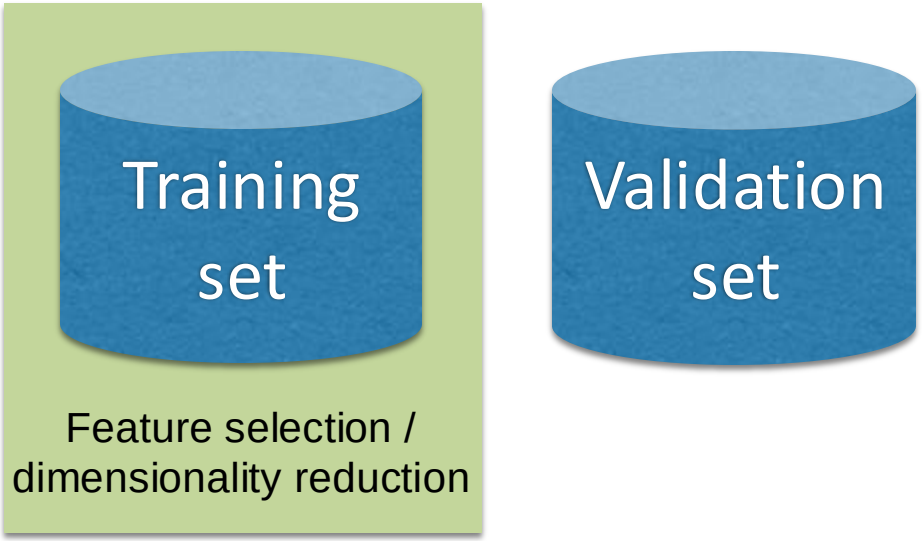
- Incorrect way: do not do feature selection (or dimensionality reduction) on the whole dataset, and then cross-validation



- Feature selection and dimensionality reduction on the whole dataset destroys cross-validation
  - reduced training set would depend on the validation set
  - thus, training is looking at the supposedly "unseen" data

Feature selection and cross-validation

- Correct way: feature selection (or dimensionality reduction) inside cross-validation, only applied to the training set



Variability

- When reporting any result, we cannot just report the mean of a test metric (e.g., k-fold cross validation, bootstrapping), we need to also report the variance
  - Better way to compare alternatives
- Statistical hypothesis testing
  - An imperfect technique applied on an imperfect setting, but very useful anyway to compare two alternatives
- Error bars
  - Another way to report metric mean and variance

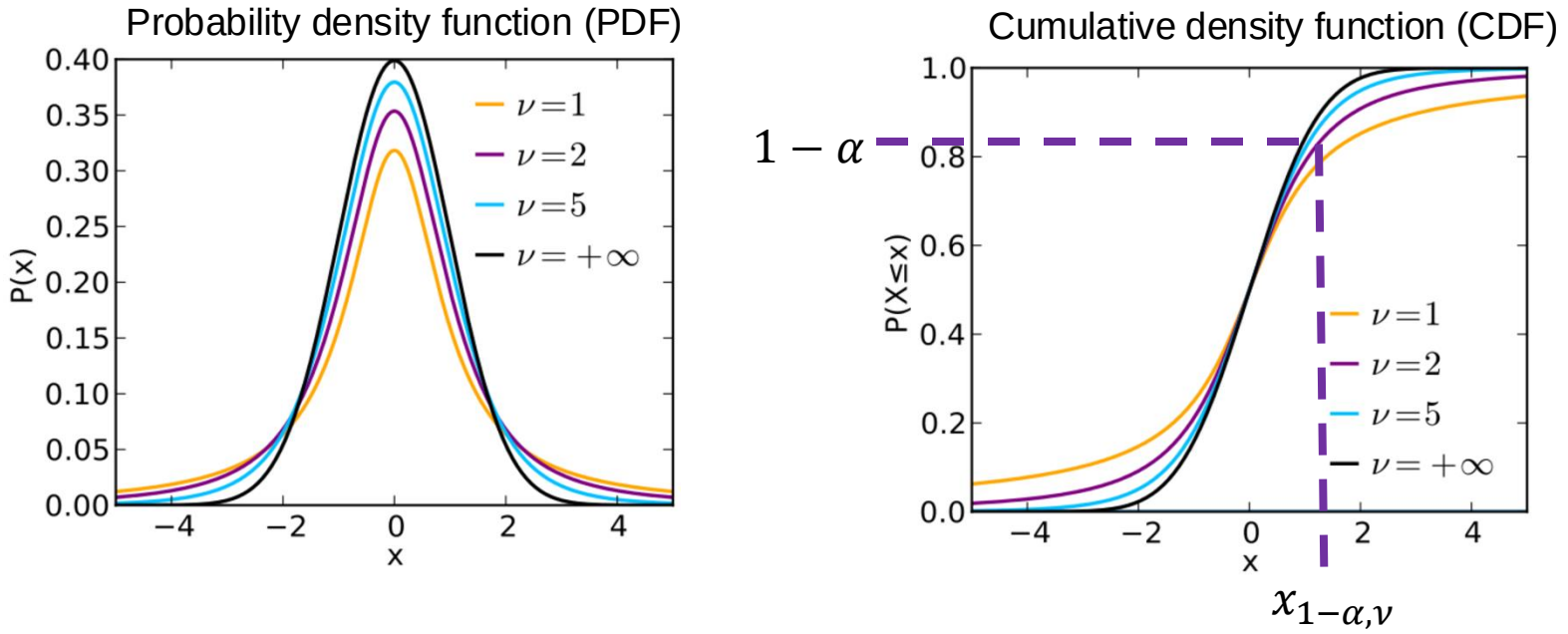
Statistical hypothesis testing

- How to compare two algorithms?
  - Not only means, also variances!
  - Here m is the number of repetitions, e.g., k for k-fold cross validation, or B for 0.632 bootstrapping
- Let $\hat{\mu}_1$ $\hat{\mu}_2$ $\hat{\sigma}_1^2$ $\hat{\sigma}_2^2$ be mean and variance of algorithms 1 and 2.
- When to reject null hypothesis $\mu_1 = \mu_2$ in favor of $\mu_1 > \mu_2$?
- Let $x = \frac{(\hat{\mu}_1 - \hat{\mu}_2)\sqrt{m}}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}}$ $\quad$ $\nu = \left\lfloor \frac{(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)^2 (m-1)}{\hat{\sigma}_1^4 + \hat{\sigma}_2^4} \right\rfloor$

Degrees of freedom of Student's t-distribution

Statistical hypothesis testing

- Student's t-distribution:



- For significance level α, degrees of freedom ν
  - Find the value $x_{1-\alpha,\nu}$ for which CDF = $1-\alpha$
- If $x > x_{1-\alpha,\nu}$ reject null hypothesis $\mu_1 = \mu_2$ in favor of $\mu_1 > \mu_2$

Hypothesis testing: example 1

- Two algorithms tested with 9-fold cross validation
- Percentage of error on each left-out fold:
  - A1: 11, 7, 13, 12, 12, 9, 10, 7, 10 $\quad \hat{\mu}_1 = 10.1 \quad \hat{\sigma}_1^2 = 4.1$
  - A2: 10, 8, 12, 10, 11, 9, 13, 7, 9 $\quad \hat{\mu}_2 = 9.9 \quad \hat{\sigma}_2^2 = 3.2$
- Can we reject null hypothesis ($\mu_1 = \mu_2$) in favor of alternate hypothesis ($\mu_1 > \mu_2$) at 5% significance level?

$$x = \frac{(\hat{\mu}_1 - \hat{\mu}_2)\sqrt{m}}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}} = \frac{(10.1 - 9.9)\sqrt{9}}{\sqrt{4.1 + 3.2}} \approx 0.22$$

$$\nu = \left\lfloor \frac{(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)^2 (m-1)}{\hat{\sigma}_1^4 + \hat{\sigma}_2^4} \right\rfloor = \left\lfloor \frac{(4.1 + 3.2)^2 8}{4.1^2 + 3.2^2} \right\rfloor \approx \lfloor 15.8 \rfloor = 16$$

- Inverse CDF $x_{1-0.05,\nu} = x_{0.95,16} = 1.75$
- $x = 0.22 \leq 1.75 = x_{0.95,16}$ then cannot reject null

Hypothesis testing: example 2

- Two algorithms tested with 9-fold cross validation
- Percentage of error on each left-out fold:
  - A1: 10, 12, 14, 13, 13, 10, 11, 10, 11
  - A2: 10, 8, 12, 10, 11, 9, 13, 7, 9

$\hat{\mu}_1 = 11.6 \quad \hat{\sigma}_1^2 = 2$

$\hat{\mu}_2 = 9.9 \quad \hat{\sigma}_2^2 = 3.2$

- Can we reject null hypothesis ($\mu_1 = \mu_2$) in favor of alternate hypothesis ($\mu_1 > \mu_2$) at 5% significance level?

$$x = \frac{(\hat{\mu}_1 - \hat{\mu}_2)\sqrt{m}}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}} = \frac{(11.6 - 9.9)\sqrt{9}}{\sqrt{2 + 3.2}} \approx 2.24$$

$$\nu = \left\lceil \frac{(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)^2 (m-1)}{\hat{\sigma}_1^4 + \hat{\sigma}_2^4} \right\rceil = \left\lceil \frac{(2 + 3.2)^2 8}{2^2 + 3.2^2} \right\rceil \approx \lceil 15.2 \rceil = 16$$

- Inverse CDF $x_{1-0.05,\nu} = x_{0.95,16} = 1.75$
- $x = 2.24 > 1.75 = x_{0.95,16}$ then reject null

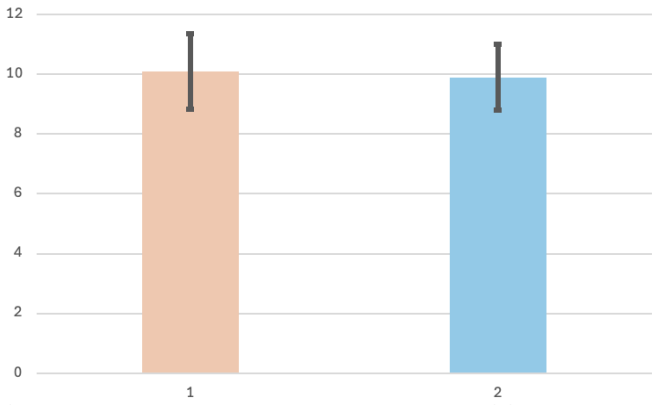
Error bars

- Two algorithms tested with 9-fold cross validation
- Percentage of error on each left-out fold:
  - A1: 11, 7, 13, 12, 12, 9, 10, 7, 10
  - A2: 10, 8, 12, 10, 11, 9, 13, 7, 9

$\hat{\mu}_1 = 10.1 \quad \hat{\sigma}_1^2 = 4.1$

$\hat{\mu}_2 = 9.9 \quad \hat{\sigma}_2^2 = 3.2$

- Error bar $\hat{\mu}_j \pm \frac{\hat{\sigma}_j}{\sqrt{m}} x_{1-\alpha/2,m-1}$
  - at 10% significance level
  - $10.1 \pm \frac{\sqrt{4.1}}{\sqrt{9}} 1.86$ which is 10.1 ± 1.26
  - $9.9 \pm \frac{\sqrt{3.2}}{\sqrt{9}} 1.86$ which is 9.9 ± 1.11



(bars overlap: no clear difference between algorithms)

What is a sample?

- In this lecture we assume that each sample is a different "unit of interest" for the experimenter
- Never sample the same "unit of interest" several times
  - In a medical application, we might be interested on knowing the accuracy (and variance) with respect to patients.
  - Taking two visits of the same patient as two different samples would be incorrect.
- Collect more data, if necessary
  - Never duplicate data as a means to claim that you have more data samples, since your data will not capture the right variability, e.g., taking 1000 pictures of 3 objects versus taking one picture, each for one of 3000 objects.