Machine learning

Laia Subirats Maté

27 de maig de 2022

Índex

1. Aprenentatge supervisat: classificació

- Nearest Neighbors
- Naïve Bayes
- Decision Trees
- Support vector machines
- Neural network models

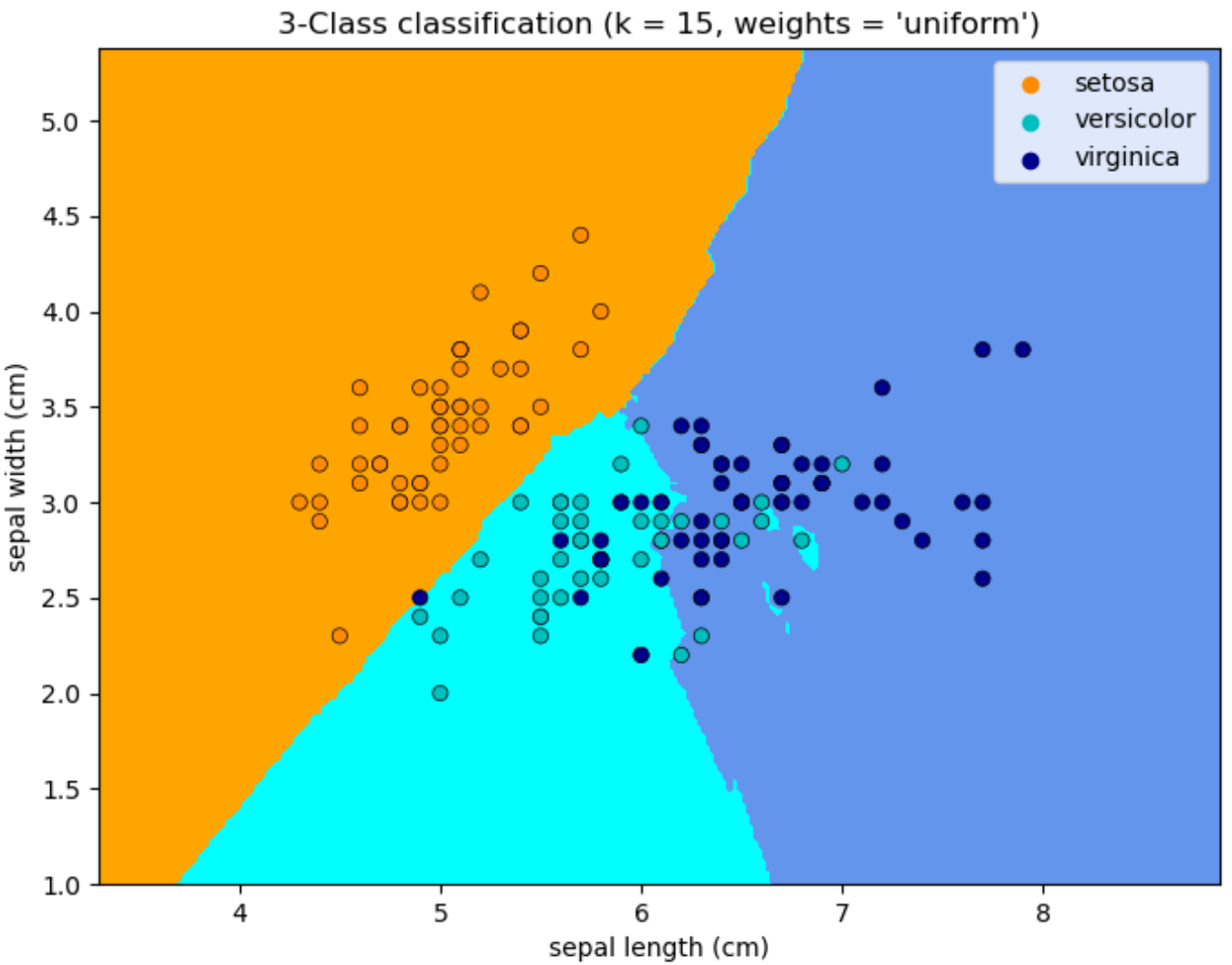
Nearest neighbors

- k: quants veïns escollim?
- La classificació basada en veïns és un tipus d'aprenentatge **basat en instàncies** o d'aprenentatge no generalitzador: no intenta construir un model intern general, sinó que simplement emmagatzema les instàncies de les dades de formació. La classificació es calcula a partir d'un **vot per majoria simple dels veïns més propers de cada punt**: a un punt de consulta se li assigna la classe de dades que té més representants dins dels veïns més propers del punt.

Nearest neighbors

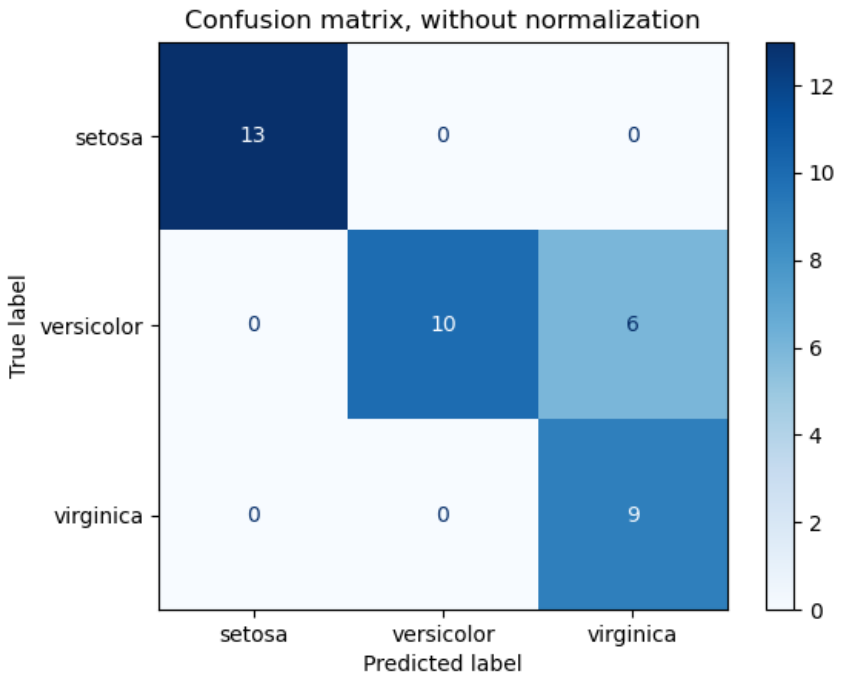
La classificació bàsica dels veïns més propers utilitza pesos uniformes: és a dir, el valor assignat a un punt de consulta es calcula a partir d'un vot de majoria simple dels veïns més propers. En algunes circumstàncies, és millor ponderar els veïns de manera que els veïns més propers contribueixin més a l'adequació. Això es pot aconseguir mitjançant la paraula clau de peses. El valor per defecte, weights = '**uniform**', assigna pesos uniformes a cada veí. weights = '**distance**' assigna pesos proporcionals a la inversa de la distància des del punt de consulta. Com a alternativa, es pot proporcionar una funció definida per l'usuari de la distància per calcular els pesos.

Nearest neighbors



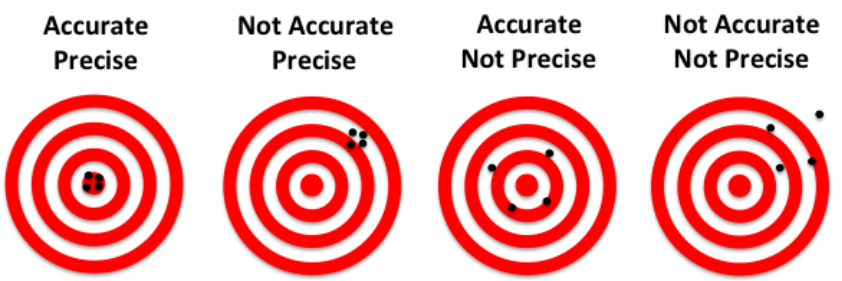
Aprenentatge automàtic

- Aprenentatge supervisat: Classificació

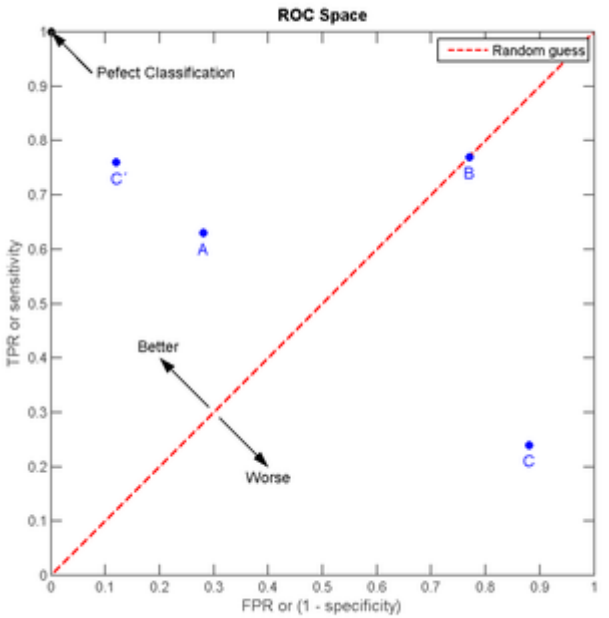


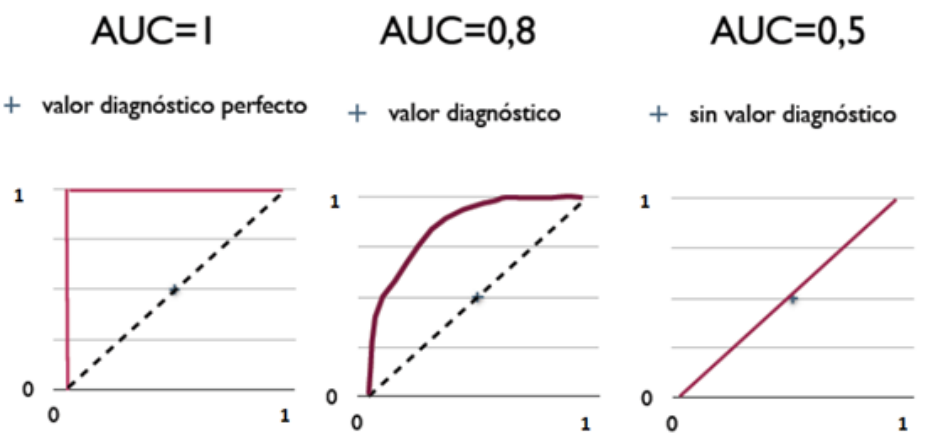
- Aprenentatge supervisat: Classificació

		Predicted Class: Positive	Predicted Class: Negative	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$



- Area under the curve (AUC – ROC → Receiver Operating Characteristic)





Naïve Bayes

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Aplicació en classificacions de documents i filtratge de **correu brossa** (spam). Requereixen una **petita quantitat de dades** de formació per estimar els paràmetres necessaris. Són extremadament **ràpids** en comparació amb mètodes més sofisticats.

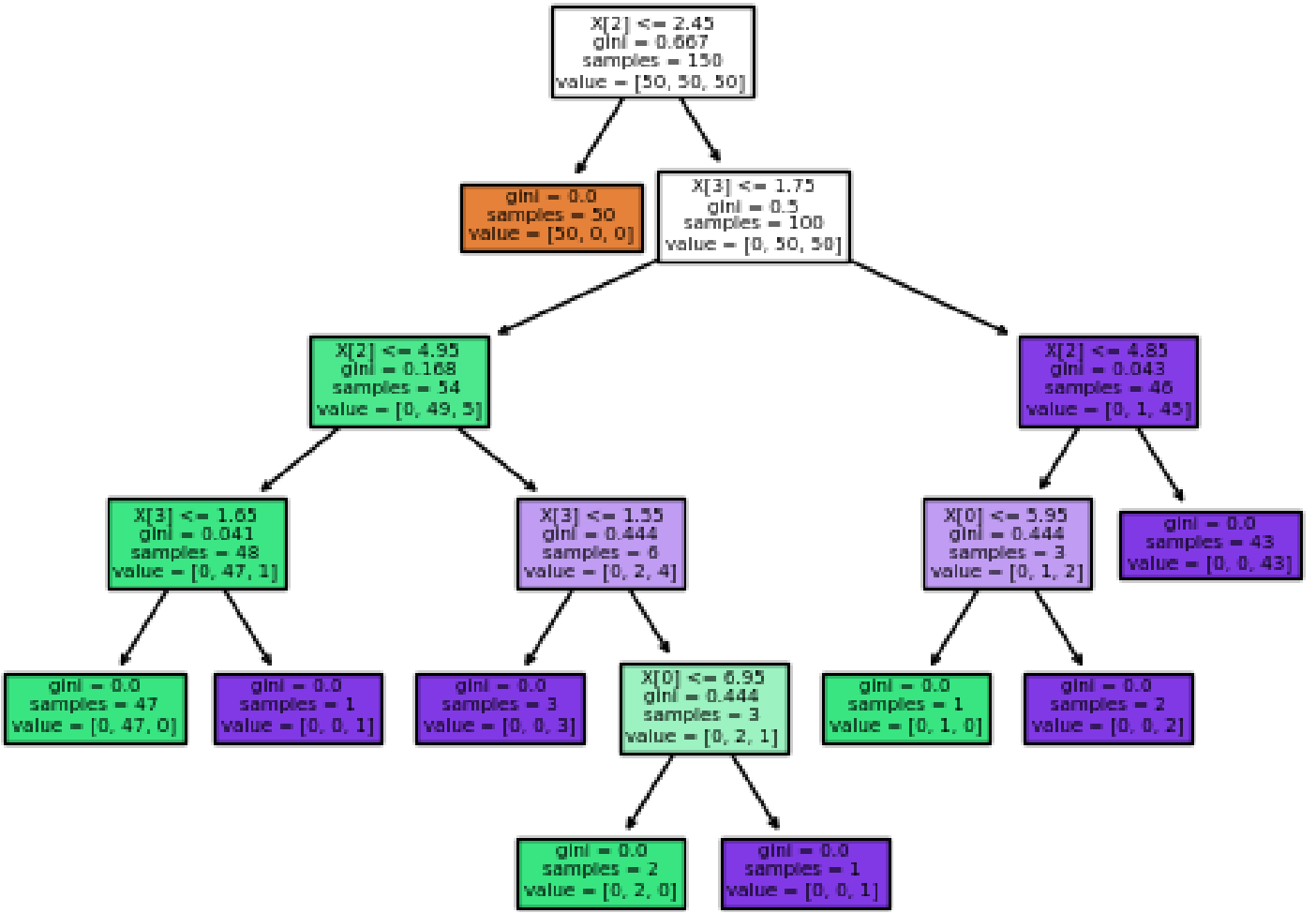
Arbres de decisió: avantatges

•Simple to understand and to **interpret**. Trees can be visualised.
•Requires **little data preparation**. Other techniques often require data normalisation, dummy variables need to be created and blank values to be removed.
• The cost of using the tree (i.e., predicting data) is **logarithmic** in the number of data points used to train the tree.
•Able to handle both **numerical and categorical data**. However scikit-learn implementation does not support categorical variables for now. Other techniques are usually specialised in analysing datasets that have only one type of variable.
•Able to handle **multi-output problems**.
•Uses a **white box model**. If a given situation is observable in a model, the explanation for the condition is easily explained by boolean logic. By contrast, in a black box model (e.g., in an artificial neural network), results may be more difficult to interpret.

Arbres de decisió: desavantatges

- Decision-tree learners can create **over-complex trees** that do not generalise the data well. This is called overfitting. Mechanisms such as pruning, setting the minimum number of samples required at a leaf node or setting the maximum depth of the tree are necessary to avoid this problem.
- Decision trees can be **unstable** because small variations in the data might result in a completely different tree being generated. This problem is mitigated by using decision trees within an ensemble.
- There are concepts that are **hard to learn** because decision trees do not express them easily, such as XOR, parity or multiplexer problems.
- Decision tree learners create **biased trees if some classes dominate**. It is therefore recommended to balance the dataset prior to fitting with the decision tree.

Arbres de decisió



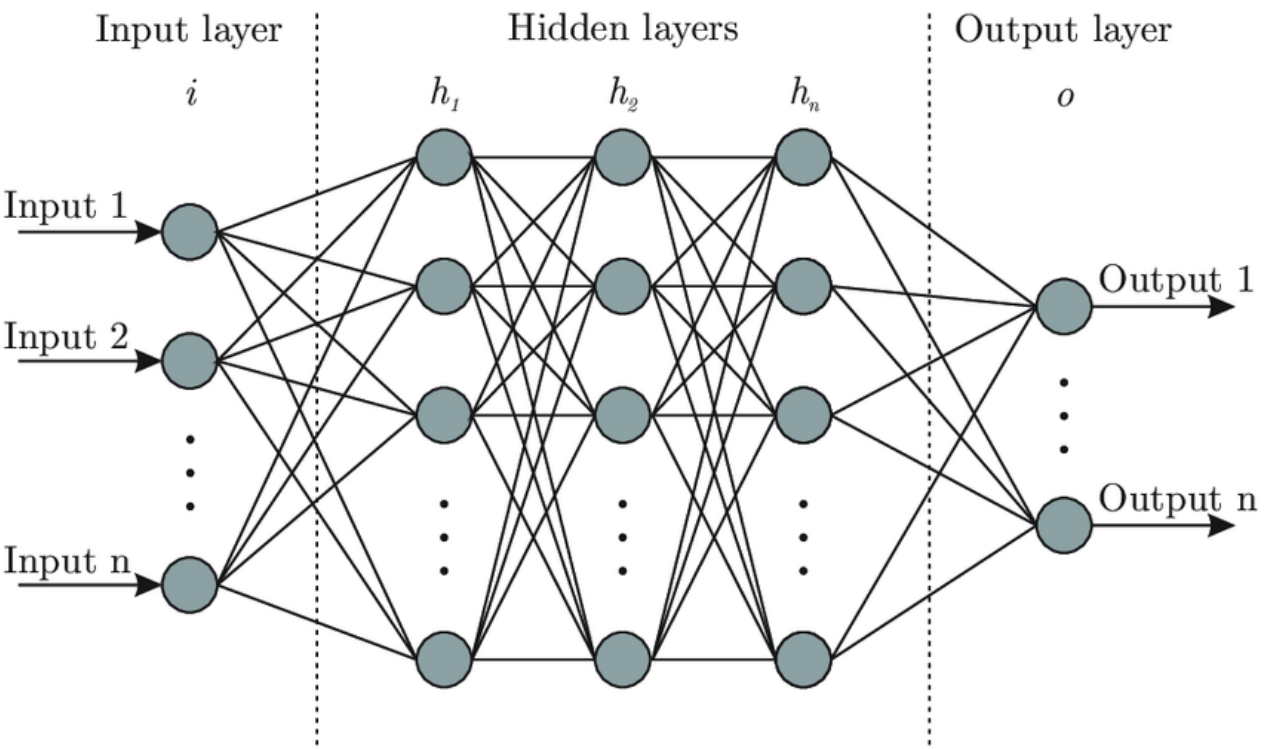
Support vector machines: avantatges

- Effective in **high dimensional spaces**.
- Still effective in cases where **number of dimensions** is greater than the **number of samples**.
- Uses a subset of training points in the decision function (called support vectors), so it is also **memory efficient**.
- **Versatile**: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

Support vector machines: desavantatges

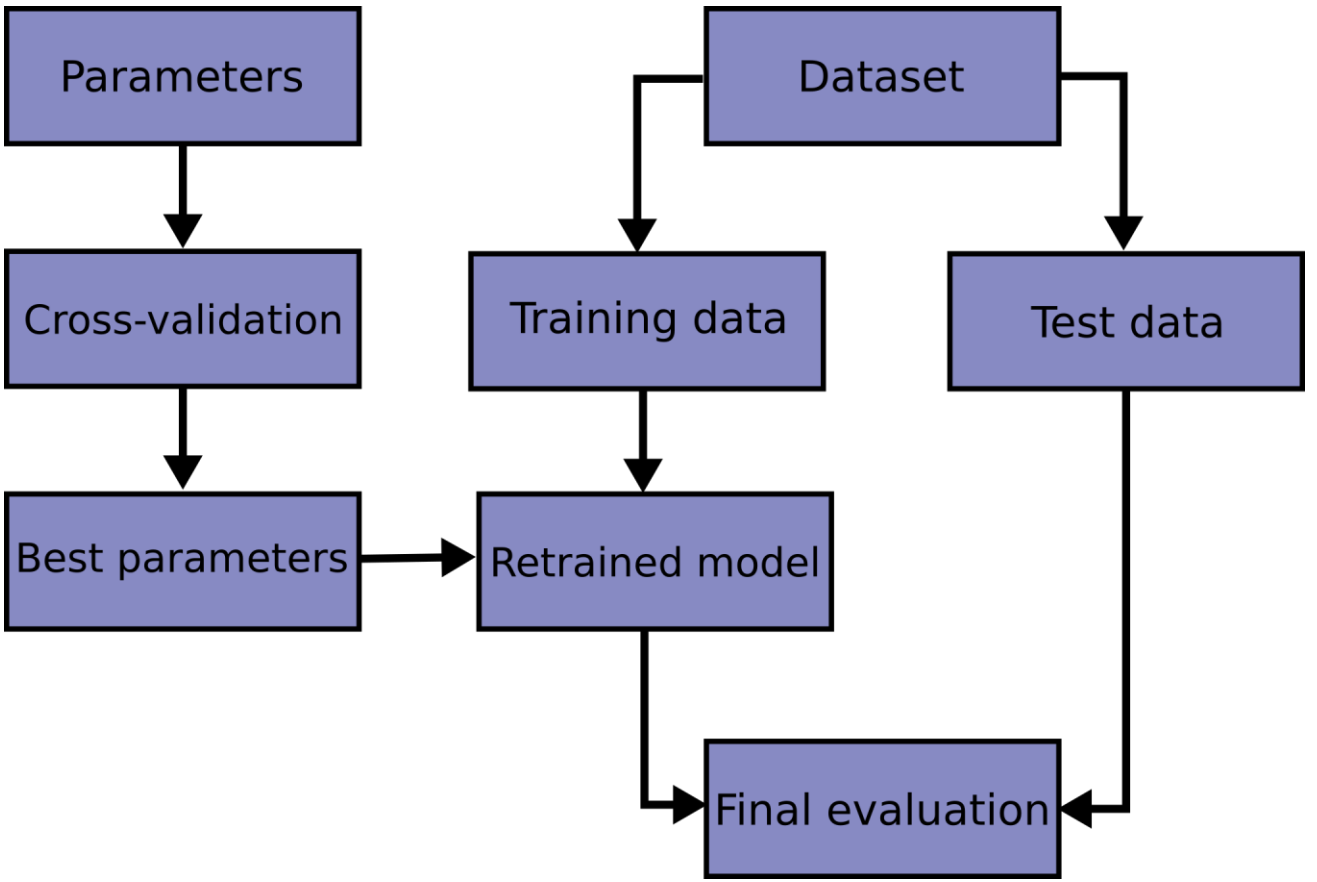
- If the number of features is much greater than the number of samples, avoid over-fitting in choosing Kernel functions and regularization term is crucial.
- SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation (see Scores and probabilities, below).

Neural networks

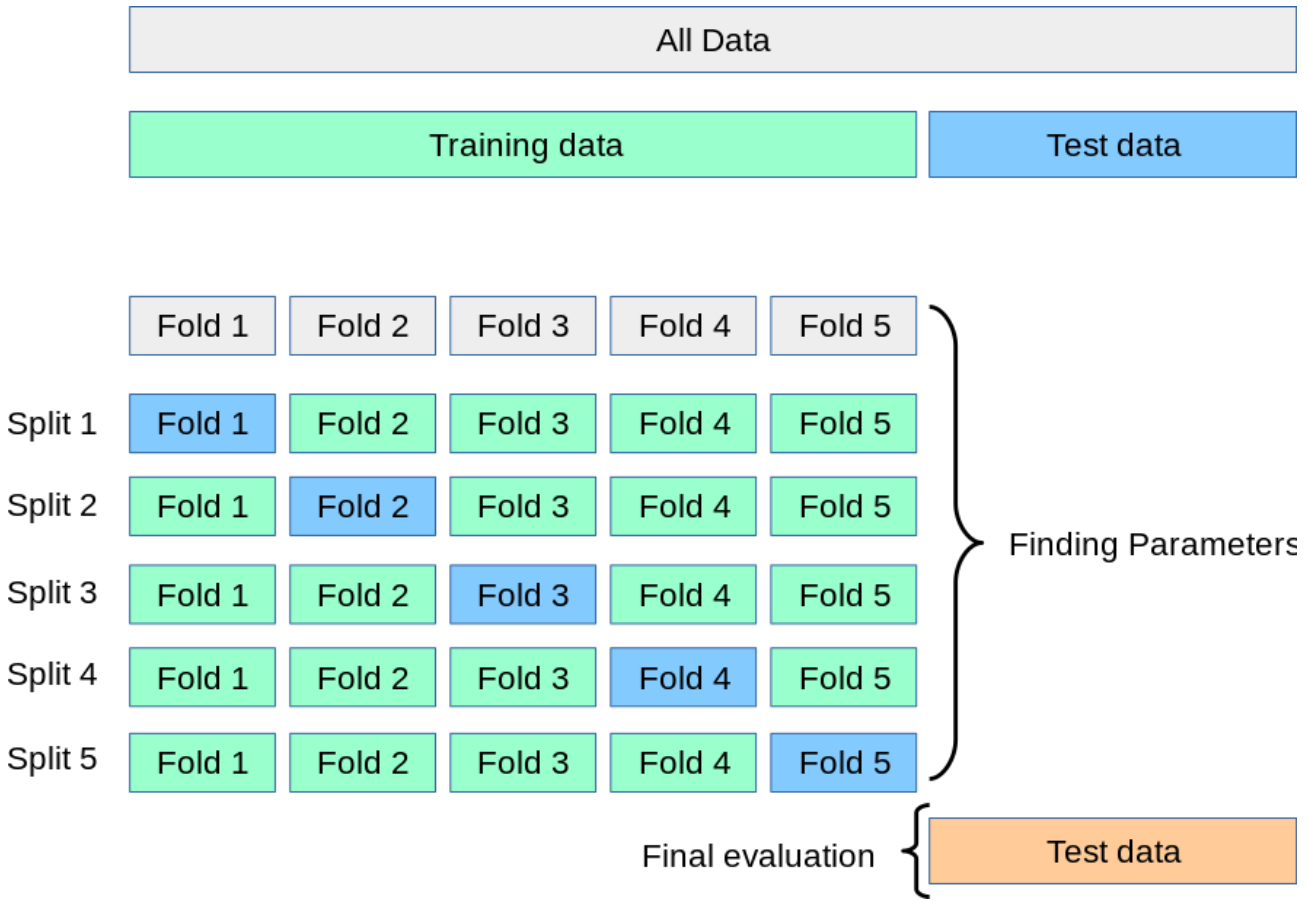


https://www.youtube.com/watch?v=bfmFfD2RIcg

Cross-validation



Cross-validation to avoid overfitting



Gràcies per la vostra col·laboració!

Laia.Subirats@eurecat.org