

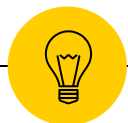


ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΤΜΗΜΑ ΦΥΣΙΚΗΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΥΠΟΛΟΓΙΣΤΙΚΗΣ ΦΥΣΙΚΗΣ



ΤΜΗΜΑ ΦΥΣΙΚΗΣ Α.Π.Θ.
Πρόγραμμα Μεταπτυχιακών Σπουδών
Υπολογιστικής Φυσικής



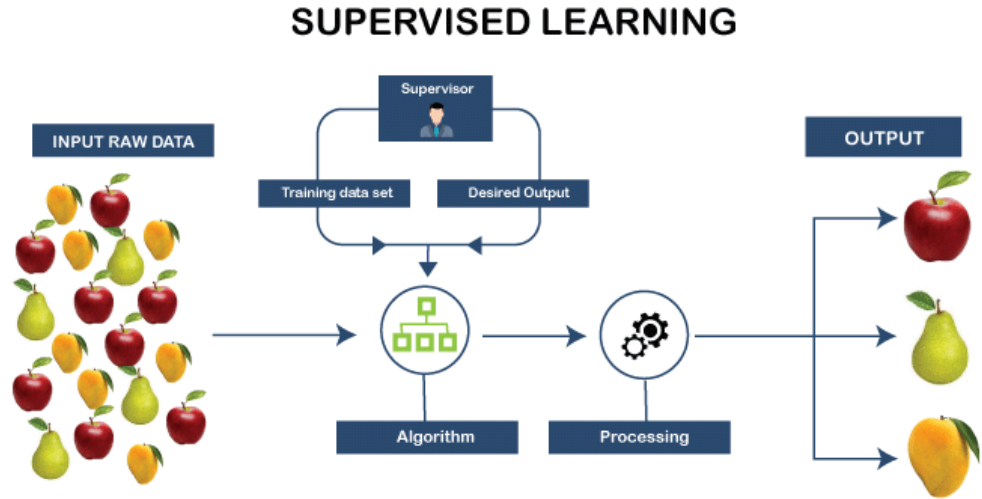
Μηχανική Μάθηση σε Python Επιβλεπόμενη Μάθηση (Supervised Learning)

Διακονίδης Θόδωρος
Ε.ΔΙ.Π

Εργαστήριο Θεωρ. Φυσικής
Τμήμα Φυσικής ΑΠΘ

Γιατί ονομάζεται έτσι; Πότε εφαρμόζεται;

- Ονομάζεται επιβλεπόμενη γιατί βασίζεται σε γνωστά δεδομένα παρατηρήσεων-αποτελεσμάτων. Ο σκοπός της είναι να κατασκευάσει ένα μοντέλο με βάση το οποίο θα μπορέσει να προβλέψει τα αποτελέσματα αγνώστων δειγμάτων.





Κατηγορίες επιβλεπόμενης μάθησης



Μπορούμε να την χωρίσουμε σε 2 κατηγορίες ανάλογα με τον τρόπο εφαρμογής:



1. Classification (κατάταξης)

Ή στην περίπτωση αυτή γίνεται κατάταξη του δείγματος σε διακριτό (district) αποτέλεσμα. Ναι/Όχι, Αλήθεια/Ψέμα, Επιβίωση/Θάνατος κλπ.



2. Regression (πρόβλεψης)

Σε αυτή την περίπτωση χρησιμοποιούνται οι αλγόριθμοι για πρόβλεψη συνεχούς αριθμητικού αποτελέσματος. Τιμή, μισθός, ηλικία κλπ.



Αλγόριθμοι εφαρμογής επιβλεπόμενης μάθησης

□ Με ποιους αλγόριθμους θα ασχοληθούμε στην Python:

1. K-Nearest Neighbors(K-NN)
2. Support Vector Machine (SVM)
3. Decision Trees
4. Random Forest



ΤΜΗΜΑ ΦΥΣΙΚΗΣ Α.Π.Θ.
Πρόγραμμα Μεταπτυχιακών Σπουδών
Υπολογιστικής Φυσικής

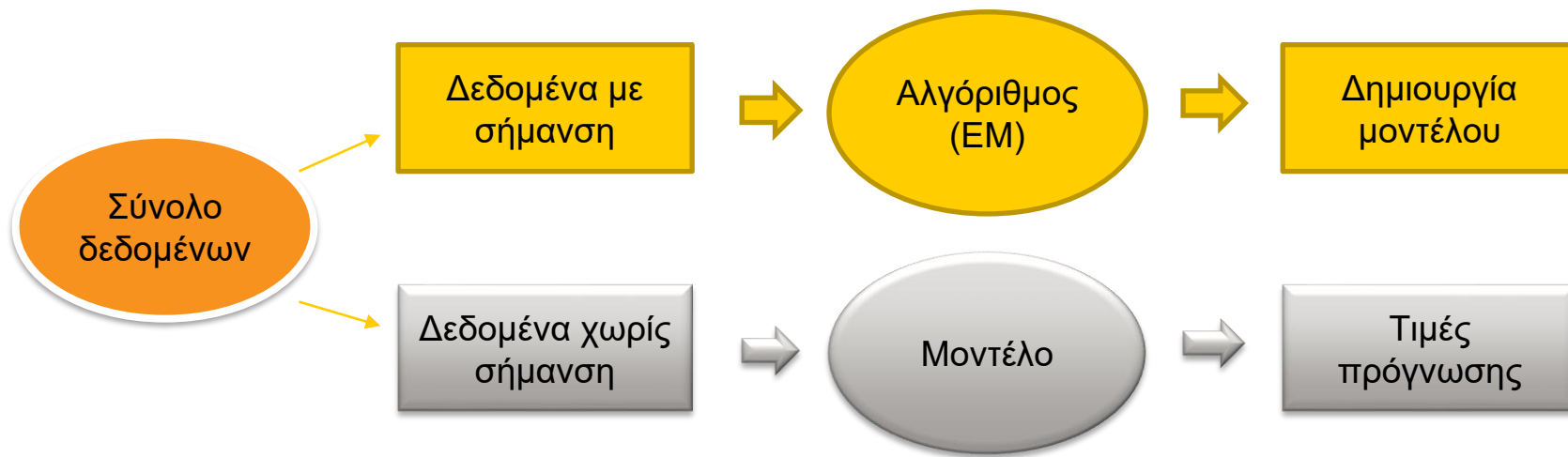


Κατάταξη (Classification)

- Η βασική χρήση των αλγορίθμων είναι η κατάταξη-πρόβλεψη ενός πλήθους δειγμάτων (**test set**) σε κάποια κατηγορία (class) με βάση τις ιδιότητες και την πληροφορία που διαθέτουν.
(Π.χ. Στη βάση δεδομένων που είδαμε, αν ο ασθενής έχει ή δεν έχει διαβήτη)
- Για να επιτευχθεί κάτι τέτοιο, απαραίτητη προϋπόθεση είναι να κατασκευαστεί ένα μοντέλο κατανεμητής (classifier) με βάση το οποίο θα γίνει η πρόβλεψη. Η εκπαίδευση (training) του υπό κατασκευή μοντέλου γίνεται με βάση ήδη γνωστά αποτελέσματα.
(Π.χ. στην περίπτωση του diabetes.csv, διαχωρίζουμε τη βάση σε 2 τμήματα. **training set** και **test set**. Επειδή γνωρίζουμε ήδη τα αποτελέσματα και για το **test set** μπορούμε να συγκρίνουμε άμεσα την επιτυχία του μοντέλου πρόβλεψης.)



Διαχωρισμός δεδομένων



Στο πάνω μέρος, το σύνολο των προς εκπαίδευση δεδομένων (**training set**) οδηγεί στην δημιουργία του μοντέλου.

Στο κάτω μέρος τα δεδομένα ελέγχου (**test set**) για τα οποία αναζητούμε πρόβλεψη (σήμανση). Με τη βοήθεια του μοντέλου επιτυγχάνεται αυτή.



Πίνακας σύγχυσης (Confusion Matrix)

- Δείχνει τον αριθμό των σωστών και λανθασμένων προβλέψεων που γίνονται για το μοντέλο ταξινόμησης που υλοποιείται σε σχέση με τα πραγματικά αποτελέσματα (τιμή στόχο) στα δεδομένα.
- Περιέχει επομένως πληροφορίες, σχετικά με την πραγματική και την προβλεπόμενη ταξινόμηση.

Two-class Binary model

		Predicted Class	
		A	B
Actual Class	A	O	X
	B	X	O

- Οι σωστές προβλέψεις βρίσκονται στην διαγώνιο του πίνακα, ενώ στα υπόλοιπα κελιά φαίνεται ο αριθμός των λανθασμένων προβλέψεων. Ιδανικά θα θέλαμε η διαγώνιος να περιέχει μεγάλους αριθμούς, ενώ τα υπόλοιπα κελιά να τείνουν όσο το δυνατόν περισσότερο στο μηδέν.



Πίνακας σύγχυσης (Confusion Matrix)

● Πίνακας Σύγχυσης για την sonar

		ΤΙΜΕΣ ΠΡΟΒΛΕΨΗΣ	
		Μ	Ρ
ΠΡΑΓΜΑΤΙΚΕΣ ΤΙΜΕΣ	Μ	ΟΡΘΗ ΠΡΟΒΛΕΨΗ Μ	ΠΡΟΒΛΕΨΗ Ρ ΕΝΩ Μ
	Ρ	ΠΡΟΒΛΕΨΗ Μ ΕΝΩ Ρ	ΟΡΘΗ ΠΡΟΒΛΕΨΗ Ρ

● Ας θεωρήσουμε ότι αναζητούμε το Ρ (Positive) και ότι η επιλογή Σ' αυτή την περίπτωση Μ είναι αρνητική (Negative).

Πίνακας σύγχυσης (Confusion Matrix)

		Actual Condition		
		FALSE	TRUE	
Predicted Condition	FALSE	TN	FN	Predicted Negative
	TRUE	FP	TP	Predicted Positive
		Actual Negative	Actual Positive	

- TN είναι ο αριθμός των αρνητικών παραδειγμάτων που έχουν ταξινομηθεί σωστά (True Negatives)
- FP είναι ο αριθμός των αρνητικών παραδειγμάτων που έχουν από λάθος ταξινομηθεί ως θετικά (False Positives)
- FN είναι ο αριθμός των θετικών παραδειγμάτων που έχουν από λάθος ταξινομηθεί ως αρνητικά (False Negatives)
- TP είναι ο αριθμός των θετικών παραδειγμάτων που έχουν ταξινομηθεί σωστά ως θετικά (True Positives).



Πίνακας σύγχυσης (Confusion Matrix)

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

$$Recall/Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F - Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

- TN είναι ο αριθμός των αρνητικών παραδειγμάτων που έχουν ταξινομηθεί σωστά (True Negatives)
- FP είναι ο αριθμός των αρνητικών παραδειγμάτων που έχουν από λάθος ταξινομηθεί ως θετικά (False Positives)
- FN είναι ο αριθμός των θετικών παραδειγμάτων που έχουν από λάθος ταξινομηθεί ως αρνητικά (False Negatives)
- TP είναι ο αριθμός των θετικών παραδειγμάτων που έχουν ταξινομηθεί σωστά ως θετικά (True Positives).

● Παλινδρόμηση Ρίζα Μέσης Τετραγωνικής Απόκλισης (RMSE)

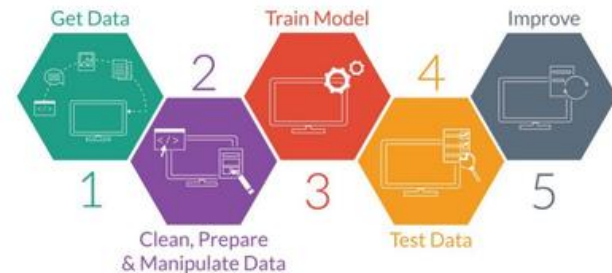
$$RMSE = \sqrt{\sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{n}}$$

- Ο βασικός δείκτης απόδοσης για ένα μοντέλο παλινδρόμησης (Regression). Μετρά τη μέση διαφορά μεταξύ των τιμών που προβλέπονται από ένα μοντέλο και των τιμών παρατήρησης. Παρέχει μια εκτίμηση του πόσο καλά το μοντέλο είναι σε θέση να προβλέψει την τιμή-στόχο (ακρίβεια)

● Διαδικασία εφαρμογής μεθόδων επιβλεπόμενης μάθησης

◎ Βήματα διαδικασίας:

1. Διαχωρισμός βάσης δεδομένων σε δεδομένα εκμάθησης (training set) και ελέγχου (test set)
2. Προετοιμασία των δεδομένων. Πιθανή χρησιμοποίηση κανονικοποίησης (normalization) ή διάφορες άλλες μεθόδους μετασχηματισμού δεδομένων.
3. Εκπαίδευση του μοντέλου με τους αλγόριθμους που προαναφέρθηκαν. Έλεγχος υπομοντελοποίησης (underfitting) και υπερμοντελοποίησης (overfitting).



● Διαδικασία εφαρμογής μεθόδων επιβλεπόμενης μάθησης

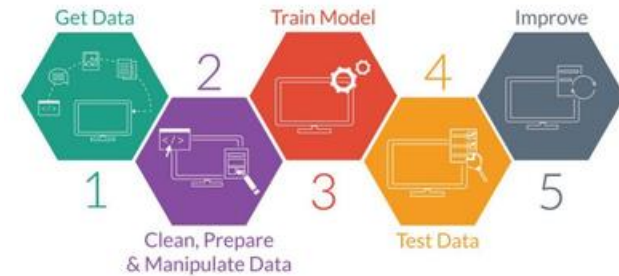
◎ Βήματα διαδικασίας (συνέχεια):

4. Εκτίμηση απόδοσης του μοντέλου.

- A) Κατασκευή πίνακα σύγχυσης (confusion matrix) (Κατάταξη)
- B) RMSE (Παλινδρόμηση)

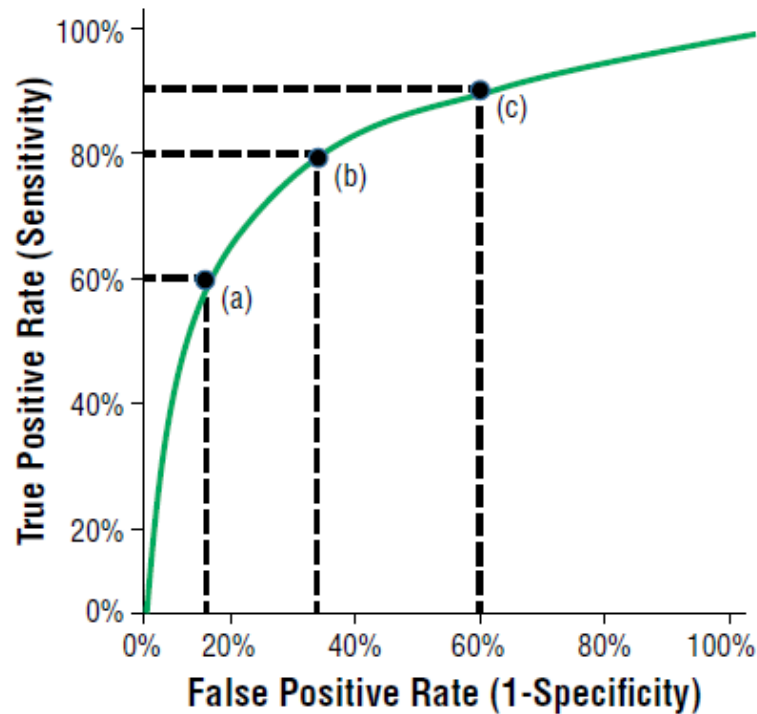
5. Ρύθμιση και επιλογή του καταλληλότερου μοντέλου

- i) Διαδικασία επικύρωσης (Cross Validation)
- ii) Έλεγχος δικτύου (Grid Search)



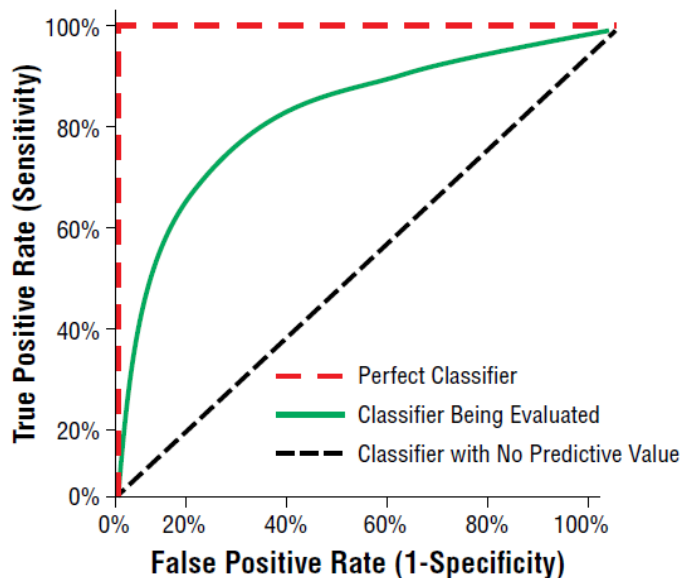
Receiver Operating Characteristic (ROC) Καμπύλη

- Αποτελεί την οπτική αναπαράσταση της σχέσης μεταξύ του πραγματικού θετικού βαθμού (True Positive Rate) ή Sensitivity και του ψευδώς θετικού (FPR) ή (1-Specificity) ενός μοντέλου για όλα τα πιθανά κατώφλια αποκοπής (cutoffs).
- Χρησιμοποιείται για να αξιολογήσει πόσο σωστά διακρίνει το μοντέλο τις θετικές και τις αρνητικές κλάσεις-ομάδες στο σύνολο δεδομένων αξιολόγησης.
- [Statquest link for ROC-AUC](#)





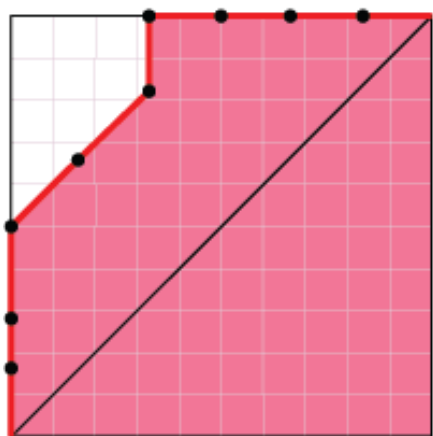
Οπτικοποίηση διαφόρων ROC



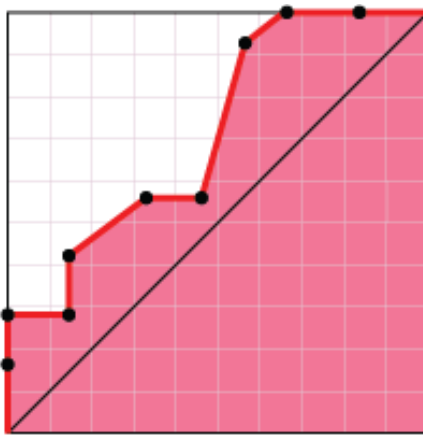
Τρεις διαφορετικοί καταναεμητές. Στην πρώτη περίπτωση (διακεκομμένη μαύρη γραμμή) ο καταναεμητής δεν έχει προβλεπτική ικανότητα (50%). Στην Τρίτη (διακεκομμένη κόκκινη) έχουμε τον ιδανικό (100%). Ένας καταναεμητής στην πράξη βρίσκεται συνήθως μεταξύ του 1^{ου} και του 3^{ου}.



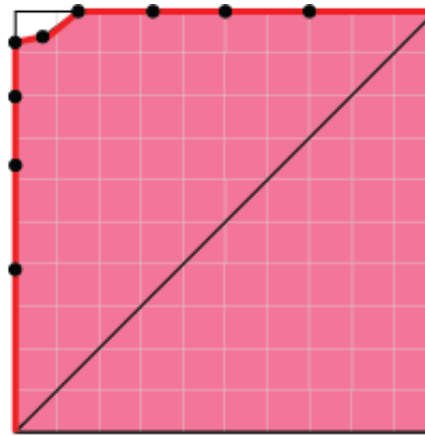
AUC (Area Under the Curve)



AUC 0.88



AUC 0.73



AUC 0.99



Είναι το μέτρο της ολικής επιφάνειας κάτω από την ROC καμπύλη. $0,5 < AUC < 1$. Είναι ένα μέτρο της δυνατότητας του εκτιμητή και του αλγορίθμου κατ' επέκταση.



Διάφοροι τύποι μετασχηματισμού δεδομένων

Αυτοί που θα χρησιμοποιήσουμε για να επεξεργαστούμε τα δικά μας δεδομένα και οι πιο διαδεδομένες είναι οι:

1. Κανονικοποίηση μεγίστου ελαχίστου (Min-max normalization)

Μετατρέπει κάθε τιμή των δεδομένων σε μια περιοχή μεταξύ 0 και 1.

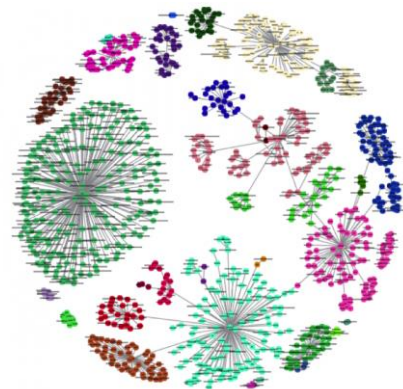
$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

2. Standardization Z-score

Μετατρέπει κάθε τιμή των δεδομένων αφαιρώντας την από τον μέσο όρο του δείγματος και διαιρώντας στη συνέχεια με την τυπική απόκλιση του δείγματος. Δεν έχει προκαθορισμένο μέγιστο ή ελάχιστο.

$$x' = \frac{x - \mu}{\sigma}$$

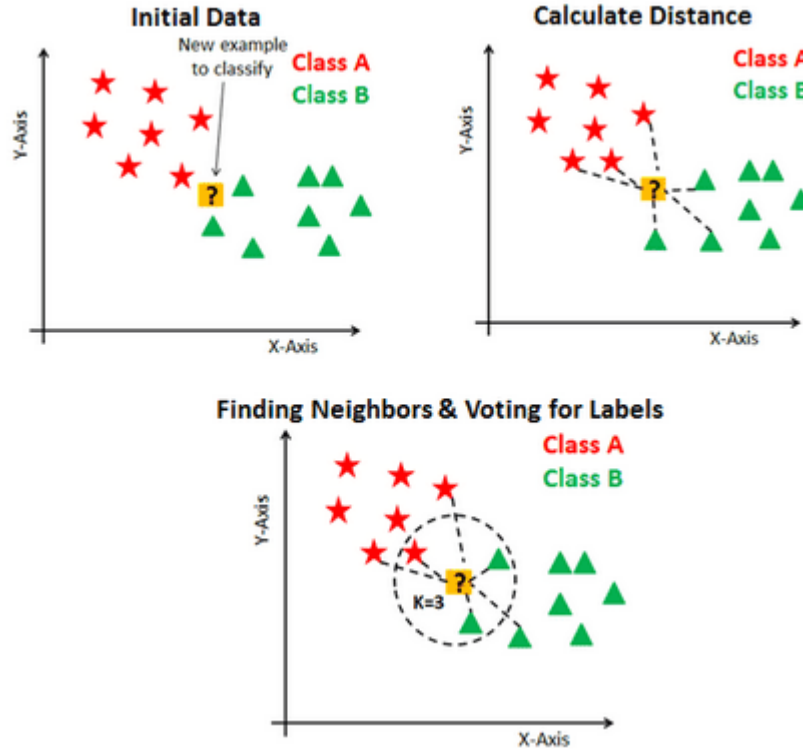
● Ο αλγόριθμος Knn (k-nearest neighbors)



- Ο απλούστερος από τους αλγορίθμους που θα εφαρμόσουμε.
- Η λογική του βασίζεται στην κατάταξη ενός καινούργιου δείγματος-παρατηρήσεως ανάλογα με το που κατατάσσονται τα k κοντινότερα δείγματα που βρίσκονται στην κοντινή περιοχή του.
- Σε γενικές γραμμές, ο knn είναι καταλληλότερος όταν οι σχέσεις μεταξύ των χαρακτηριστικών και των τάξεων-στόχων είναι πολυάριθμες, περίπλοκες ή εξαιρετικά δύσκολες προς κατανόηση αλλά τα στοιχεία παρόμοιου τύπου τάξης τείνουν να είναι αρκετά ομοιογενή.
- Από την άλλη αν τα δεδομένα μας έχουν θόρυβο ή δεν υπάρχει ξεκάθαρος διαχωρισμός τα πράγματα γίνονται δύσκολα για τον knn .



Οπτικοποίηση των σταδίων του knn



- Ένα καινούργιο στοιχείο που ανήκει στο test set επιλέγεται που θα καταλήξει με voting των κοντινότερων σημείων ως προς αυτό δειγμάτων του training set.
- Η διαδικασία επαναλαμβάνεται μέχρι να κατοχυρωθούν όλα τα δείγματα του test set σε αντίστοιχα αποτελέσματα.

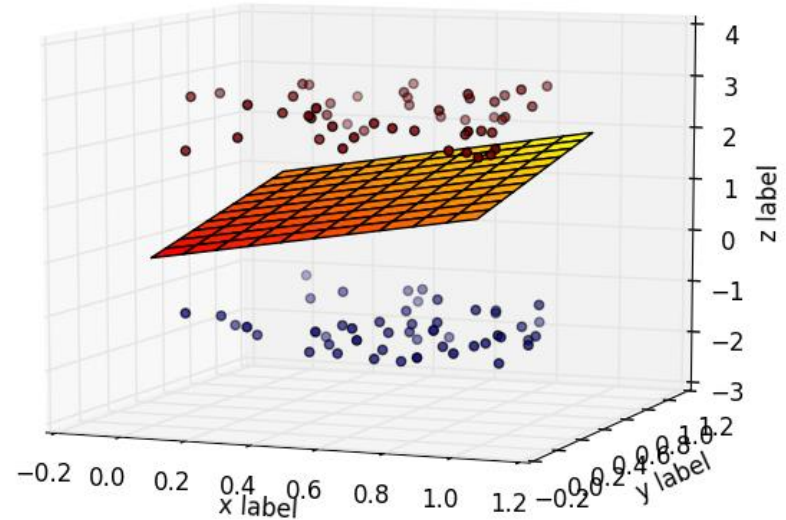
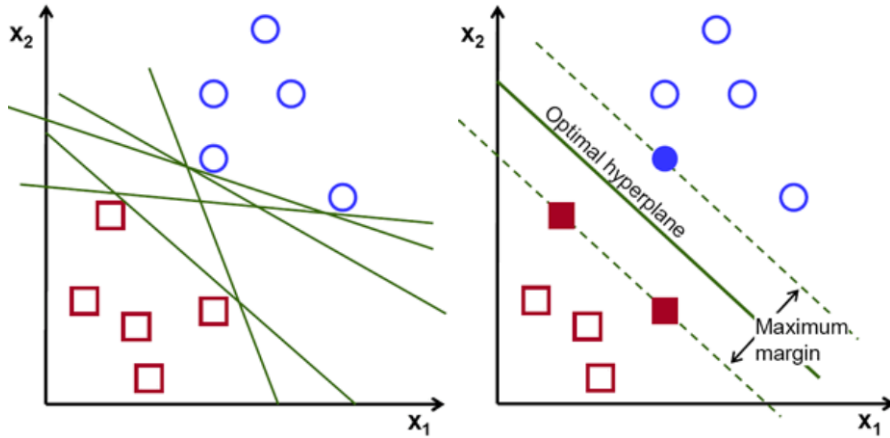


Ο αλγόριθμος SVM (Support Vector Machine)

- Ο στόχος του αλγόριθμου SVM είναι να σχηματίσει ένα τοίχος-όριο (boundary) το οποίο και ονομάζεται υπερεπίπεδο (hyperplane) και το οποίο διαχωρίζει το χώρο και δημιουργεί ομογενή τμήματα-ομάδες στην κάθε πλευρά.
- Μπορούμε να πούμε ότι συνδυάζει τις ιδιότητες του k-NN αλγορίθμου με αυτές της γραμμικής παλινδρόμησης.
- Αυτό του επιτρέπει να μοντελοποιεί πολύ πολύπλοκες σχέσεις μεταξύ των δεδομένων.

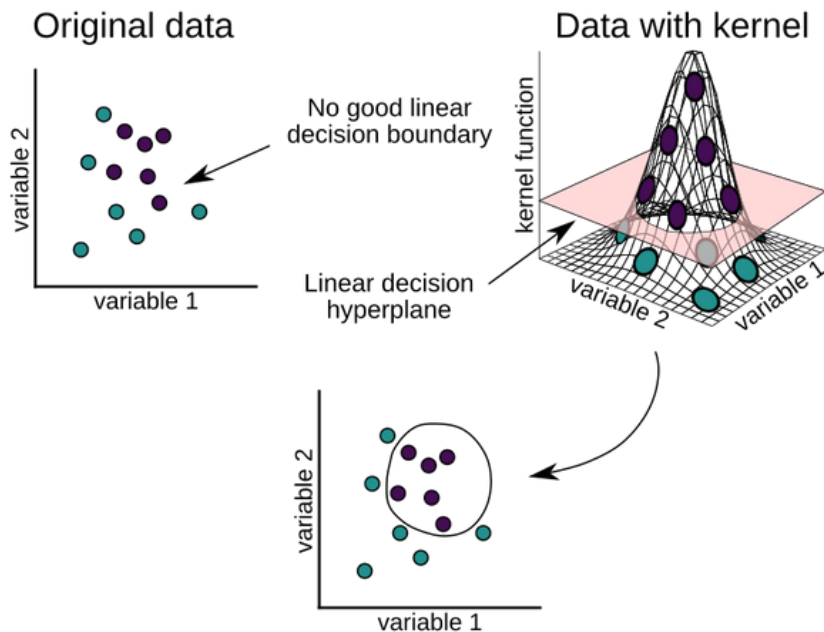


Οπτικοποίηση των σταδίων του SVM αλγορίθμου (δεδομένα LS)



- ☉ Όταν τα δεδομένα μας είναι γραμμικώς διαχωρίσιμα (linearly separable) ο αλγόριθμος svm βρίσκει την καταλληλότερη διαχωριστική γραμμή (επίπεδο σε 3 διαστάσεις) για να κατανείμει τα δείγματά μας.

Οπτικοποίηση των σταδίων του SVM αλγορίθμου δεδομένα (NLS)



Όταν τα δεδομένα μας είναι μη γραμμικώς διαχωρίσιμα (**non linearly separable**) τότε επιλέγοντας την καταλληλότερη καμπύλη (kernel) μπορούμε να τα διαχωρίσουμε κατάλληλα. (**Kernel SVM αλγόριθμος ή αλλιώς the kernel trick**)

Υπάρχουν διάφορα kernels ανάλογα με τη φύση των δεδομένων *radial (gaussian)*, *polynomial*, *sigmoid*.



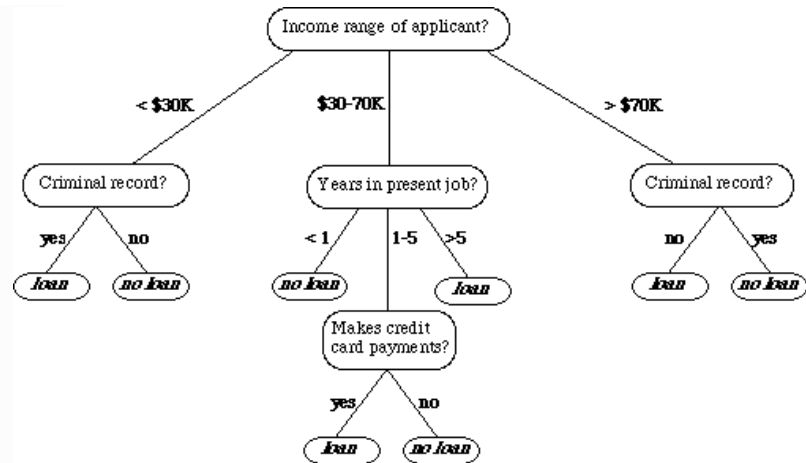
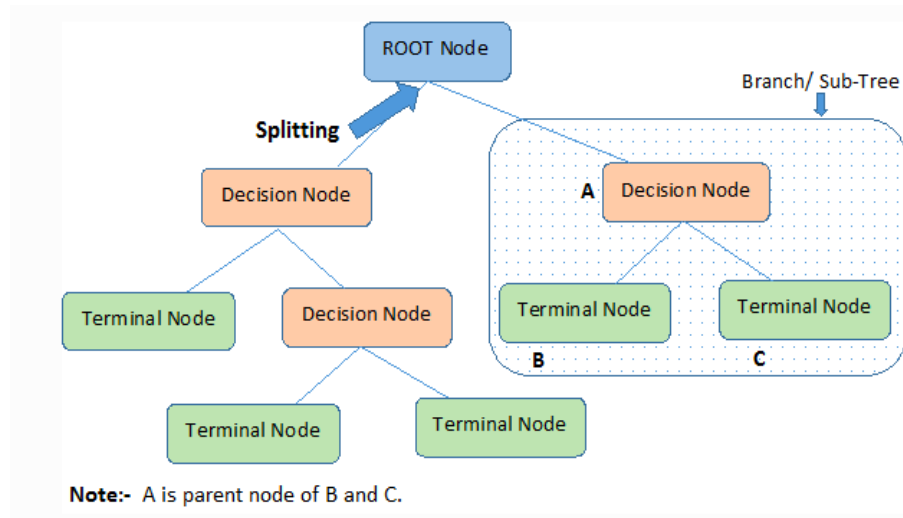
Δέντρα Απόφασης (Decision trees)



- Τα δένδρα απόφασης χρησιμοποιεί μια λογική δομή που μοιάζει με δέντρο για να αναπαραστήσει τη σχέση μεταξύ των προβλεπτών (predictors) και του αποτελέσματος (target).
- Τα δέντρα αποφάσεων κατασκευάζονται με βάση μια προσέγγιση διαίρει και βασίλευε, όπου το αρχικό σύνολο δεδομένων (root) χωρίζεται συνεχώς σε μικρότερα υποσύνολα (branches) μέχρι κάθε υποσύνολο να είναι όσο το δυνατόν πιο ομοιογενές.
- Οι κόμβοι τέλους του δέντρου είναι γνωστοί ως κόμβοι φύλλων (leaves). Αυτοί οι κόμβοι αντιπροσωπεύουν το προβλεπόμενο αποτέλεσμα με βάση το σύνολο των αποφάσεων που λαμβάνονται από τον ριζικό κόμβο, μέσω των κόμβων απόφασης (decision nodes) στον κόμβο φύλλων (terminal nodes).



Οπτικοποίηση των σταδίων των δέντρων απόφασης



- Η διαδικασία διαίρει και βασίλευε (divide and conquer) σε εφαρμογή στο δέντρο του αλγόριθμου των δέντρων απόφασης. Ο διαχωρισμός σε κόμβους απόφασης τα λεγόμενα και κλαδιά (branches) του δέντρου και η τελικοί κόμβοι της απόφασης.



Ο αλγόριθμος Random Forest



Ensemble methods

Μέθοδοι που χρησιμοποιούν ένα πλήθος από μοντέλα για την κατασκευή ενός ικανού και αποτελεσματικού μοντέλου.



Bagging (Bootstrap Aggregating)

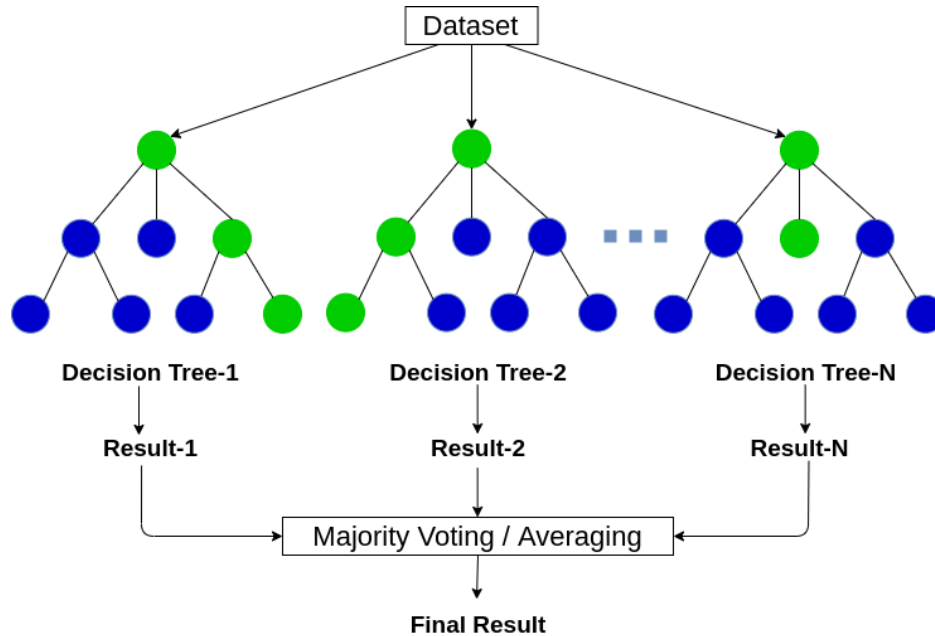
Μία από τις πιο κοινές ensemble μεθόδους. Αποτελείται από διάφορα μοντέλα που εκπαιδεύονται παράλληλα. Από κάθε μοντέλο προκύπτει μια πρόβλεψη. Αυτή που ανήκει στην πλειοψηφία λαμβάνεται ως η οριστική (hard voting). Εναλλακτικά υπολογίζονται πιθανότητες από κάθε μοντέλο και το αποτέλεσμα της πρόβλεψης αυτού με την μεγαλύτερη, επιλέγεται (soft voting).




Η **Random Forest** είναι από τις πιο δημοφιλής bagging μεθόδους. Παίρνει το όνομα από το γεγονός ότι αποτελείται από ένα μεγάλο πλήθος decision tree προβλεπτών αποτελώντας ένα δάσος (forest).



Οπτικοποίηση των σταδίων του Random Forest



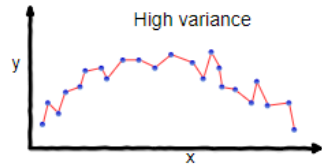
Ένα σύνολο από decision tree αλγορίθμους χρησιμοποιείται σε συνδυασμό για τον υπολογισμό του τελικού αποτελέσματος. Εδώ με βάση την πλειοψηφία (hard voting)



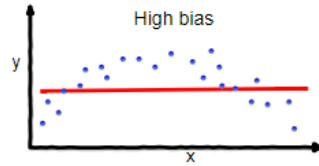
Η ανταλλαξιμότητα μεταξύ προκατάληψης και διακύμανσης (The bias-variance trade-off)

- **Προκατάληψη (Bias)** : Είναι το εγγενές σφάλμα που προκύπτει από τον ταξινομητή (classifier) ακόμη και με άπειρα δεδομένα εκπαίδευσης. Ο ταξινομητής είναι "προκατειλημμένος" σε ένα συγκεκριμένο είδος λύσης (π.χ. γραμμική παλινδρόμηση). Με άλλα λόγια, η προκατάληψη είναι εγγενής στο μοντέλο.
- **Διακύμανση (Variance)**: Καταγράφει πόσο αλλάζει ο ταξινομητής σας εάν εκπαιδεύσετε το μοντέλο σας σε διαφορετικό σύνολο εκπαίδευσης. Πόσο "υπερβολικά εξειδικευμένος" είναι ο ταξινομητής σας σε ένα συγκεκριμένο σύνολο εκπαίδευσης (overfitting).
- Αυτό που προσπαθείται πάντα να επιτευχθεί είναι μια ισορροπία μεταξύ των δύο, έτσι ώστε το ιδανικό μοντέλο να εμφανίζει το μικρότερο δυνατό σφάλμα.

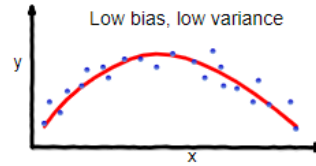
Η ανταλλαξιμότητα μεταξύ προκατάληψης και διακύμανσης (The bias-variance trade-off)



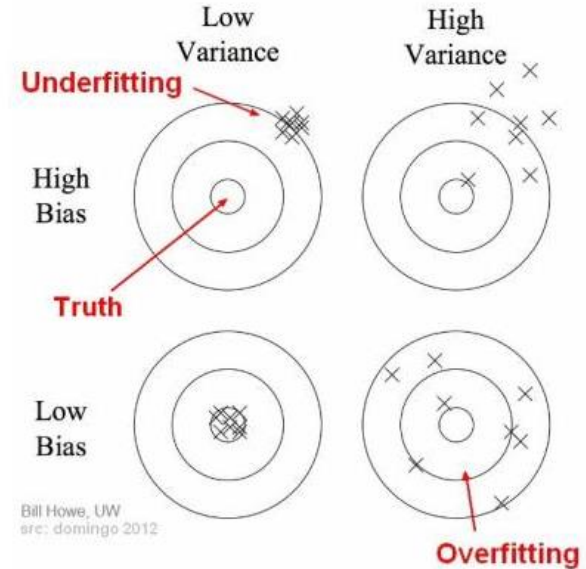
overfitting



underfitting



Good balance



- ☉ (Αριστερό σχήμα) Ακραίες επιλογές μοντέλων και μοντέλο με ιδανική ισορροπία.
- ☉ (Δεξί σχήμα) Γραφική αναπαράσταση του bias-variance trade-off.



Διαδικασία Επικύρωσης (Cross Validation)

- Η πραγματική αξία ενός μοντέλου έγκειται στο γεγονός της ορθής πρόβλεψης αγνώστων δεδομένων-παρατηρήσεων του σε δεδομένα πραγματικού κόσμου (real world examples).
- Αυτό που είναι σημαντικό είναι το μοντέλο να φτιάχνεται σωστά έτσι ώστε να ενσωματώνει γενικούς κανόνες ισχύος και όχι εξειδικευμένους, που μπορούν να ισχύουν για κάποιο συγκεκριμένο σύνολο εκπαίδευσης (train model) και που οδηγούν σε προβλήματα π.χ. υπερπροσαρμογής (overfitting).
- Υπάρχουν διάφοροι μέθοδοι-διαδικασίες επικύρωσης της ικανότητάς των μοντέλων μπορούμε να τους χωρίσουμε σε δυο κατηγορίες. **Exhaustive** και **non exhaustive**.



Τι ακριβώς κάνουν;

- ☉ Τα βήματα σε μια τέτοια διαδικασία είναι τα εξής:
 1. Παίρνουμε δείγμα από τα δεδομένα μας που γνωρίζουμε τα αποτελέσματα. (τμήμα του train set)
 2. Εκπαιδεύουμε το μοντέλο χρησιμοποιώντας τα υπόλοιπα δεδομένα. (τα εναπομείναντα δεδομένα του train set)
 3. Χρησιμοποιούμε το μοντέλο που έχουμε αποκόψει για τον έλεγχο.



Διαδικασίες επικύρωσης

☉ Exhaustive

1. Μέθοδος «άφησε ένα έξω»

(leave one out)

Αφήνεις μια παρατήρηση μόνο έξω από το σύνολο και χτίζεις το μοντέλο με τις υπόλοιπες.

Πλεονέκτημα: Χρησιμοποιούνται όλα τα δεδομένα.

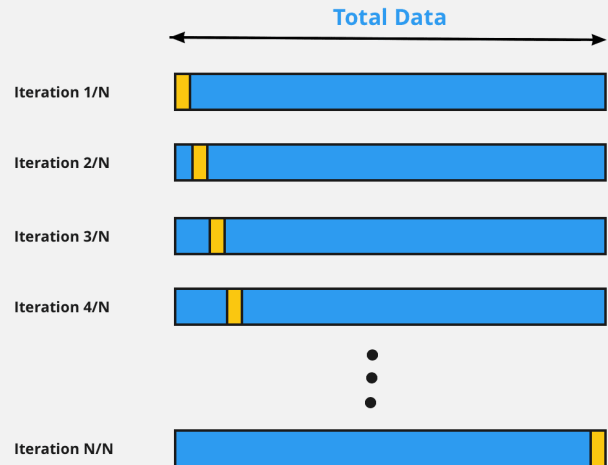
Μειονέκτημα: Μεγάλος χρόνος διαδικασίας

2. Μέθοδος «άφησε p έξω»

(leave p out)

Παραλλαγή του πρώτου, πιο μεγάλος χρόνος διαδικασίας. Αντί για n περιπτώσεις έχουμε όλες τα πιθανά γκρουπ p παρατηρήσεων από το σύνολο των n δειγμάτων.

LOOCV: Leave One Out Cross Validation



dataaspirant.com



Διαδικασίες επικύρωσης

☉ Non Exhaustive

1. Μέθοδος Holdout

Το σύνολο δεδομένων διασπάται σε δύο υποσύνολα, (αφού πρώτα γίνει ανακάτωμα). Το **σύνολο εκπαίδευσης** (training set) και το **σύνολο επικύρωσης**.

Πλεονέκτημα: Γρήγορη διαδικασία.

Μειονέκτημα: Δεν είμαστε σίγουροι για την αντιπροσωπευτικότητα του μοντέλου.

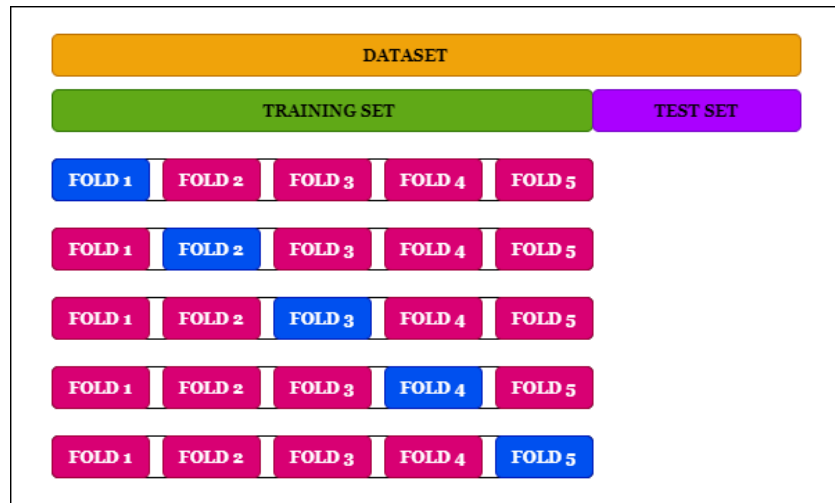
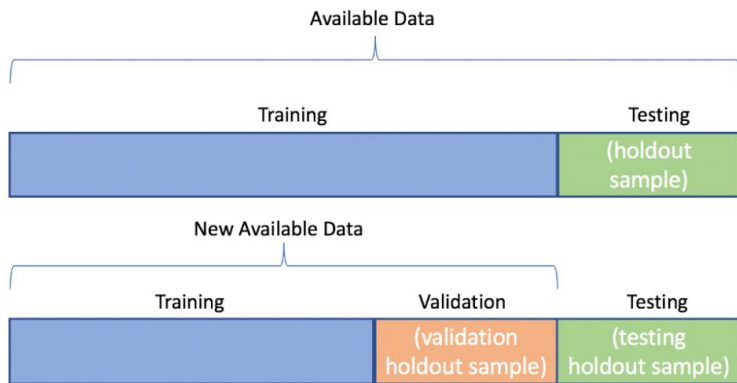
2. Μέθοδος k-fold

Η πιο δημοφιλής. Το σύνολο δεδομένων διασπάται σε k υποσύνολα. Κάθε υποσύνολο περιέχει διαφορετικές παρατηρήσεις. Η επιλογή των υποσυνόλων είναι τυχαία. Ένα από τα υποσύνολα χρησιμοποιείται ως σύνολο επικύρωσης και τα υπόλοιπα το σύνολο εκπαίδευσης. Το μοντέλο εκπαιδεύεται χρησιμοποιώντας το σύνολο εκπαίδευσης και δοκιμάζεται έναντι του συνόλου επικύρωσης. Η διαδικασία επαναλαμβάνεται k φορές, κάθε φορά χρησιμοποιώντας ένα διαφορετικό σύνολο ως σύνολο επικύρωσης και τα υπόλοιπα $k-1$ ως σύνολο εκπαίδευσης. Στο τέλος υπολογίζεται η μέση επίδοση του μοντέλου.



Απεικόνιση των Non-exhaustive μεθόδων

Cross-validation



Μέθοδος Holdout και k-fold cross validation. Στην πρώτη περίπτωση έχουμε δύο υποσύνολα του training set. Στην άλλη επιλέγουμε κ τυχαία (που συνολικά συμπληρώνουμε το data set)



Έλεγχος παραμέτρων (Grid Search) για επιπλέον ρύθμιση του μοντέλου

- Σε αρκετά από τα μοντέλα που εξετάζουμε υπάρχουν οι λεγόμενοι *hyperparameters* ενός μοντέλου, παράμετροι που είναι εξωτερικοί ως προς το μοντέλο και των οποίων η τιμή δεν μπορεί να εκτιμηθεί από τα δεδομένα. Π.χ. η τιμή *k* στον *knn* αλγόριθμο, ο αριθμός των δένδρων απόφασης **number of trees** στον *randomforest* κλπ.
- Η τιμή του υπερπαραμέτρου πρέπει να οριστεί πριν από την έναρξη της μαθησιακής διαδικασίας.
- Ο έλεγχος μιας ομάδας αυτών, βοηθά στην επιλογή των καταλληλότερων μοντέλων για ακόμη πιο ακριβείς προβλέψεις.