Αναγνώριση Προτύπων - Μηχανική Μάθηση

Επιλογή μοντέλου

Γιάννης Παναγάκης

Workflow για την Αναγνώριση Προτύπων

1. Διατύπωση του προβλήματος ως πρόβλημα μηχανικής μάθησης

2. Συλλογή δεδομένων

3. Προεπεξεργασία δεδομένων

4. Επιλογή του στατιστικού μοντέλου που θα χρησιμοποιηθεί

5. Εκπαίδευση (learn/train/estimate/fit…) του στατιστικού μοντέλου χρησιμοποιόντας δεδομένα εκπαίδευσης

6. Αξιολόγηση της επίδοσης του μοντέλου

Workflow για την Αναγνώριση Προτύπων

1. Διατύπωση του προβλήματος ως πρόβλημα μηχανικής μάθησης
2. Συλλογή δεδομένων
3. Προεπεξεργασία δεδομένων
4. Επιλογή του στατιστικού μοντέλου που θα χρησιμοποιηθεί
5. Εκπαίδευση (learn/train/estimate/fit…) του στατιστικού μοντέλου χρησιμοποιώντας δεδομένα εκπαίδευσης
6. Αξιολόγηση της επίδοσης του μοντέλου

Επιλογή μοντέλου

— Επιλογή μοντέλου: Εκτίμηση της επίδοσης διαφορετικών μοντέλων μάθησης ώστε να επιλέξουμε το πιο αποδοτικό.

— Ο βασικός στόχος των αλγορίθμων μηχανικής μάθησης είναι η επίδοση τους να είναι καλή σε νέα δεδομένα εισόδου τα οποία **δεν** περιλαμβάνονται στο σύνολο εκπαίδευσης. Αυτά τα νέα δεδομένα αποτελούν το σύνολο ελέγχου (test set).

— Η ιδιότητα των αλγορίθμων μηχανικής μάθησης να είναι αποδοτικοί σε νέα δεδομένα εκτός συνόλου εκπαίδευσης, δηλαδή στο σύνολο ελέγχου, ονομάζεται γενίκευση (generalization).

Q: Πώς μπορούμε να γνωρίζουμε την επίδοση του αλγορίθμου στο σύνολο ελέγχου όταν έχουμε στη διάθεση μας μόνο κάποιο σύνολο δεδομένων εκπαίδευσης;

Διαδικασία παραγωγής δεδομένων

— [Βασική υπόθεση στη στατιστική μάθηση: identically and independently distributed (i.i.d) data] Τα δεδομένα εκπαίδευσης και ελέγχου προέρχονται από την ίδια άγνωστη αλλά σταθερή κατανομή $p(x,y)$. Επιπλέον, όλα τα δεδομένα παράγονται ανεξάρτητα μεταξύ τους από τη $p(x,y)$.

—Η κατανομή $p(x,y)$ μοντελοποιεί διαφορετικές πηγές αβεβαιότητας (π.χ. αβεβαιότητα ως προς το χώρο υποθέσεων ή θόρυβος στα δεδομένα).

Q: Πώς μπορούμε να γνωρίζουμε την επίδοση του αλγορίθμου στο σύνολο ελέγχου όταν έχουμε στη διάθεση μας μόνο κάποιο σύνολο δεδομένων εκπαίδευσης;

A: Η i.i.d. υπόθεση επί της διαδικασίας παραγωγής των δεδομένων (εκπαίδευσης και ελέγχου) μας επιτρέπουν να μελετήσουμε τη σχέση μεταξύ σφάλματος εκπαίδευσης και ελέγχου.

Έννοιες Στατιστικής Μάθησης (1/2)

- Σύνολο εκπαίδευσης: $\mathscr{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N} \sim p$

- Το πρόβλημα μάθησης έγκειται στο να μάθουμε μια παραμετρική συνάρτηση (στατιστικό μοντέλο): $f : \mathscr{X} \to \mathscr{Y}$

- Η συνάρτηση που μαθαίνουμε ονομάζεται υπόθεση: $f \in \mathscr{H}$

- Ο χώρος υποθέσεων $\mathscr{H}$ περιλαμβάνει όλες τις συναρτήσεις που παράγονται από το μοντέλο.

Έννοιες Στατιστικής Μάθησης (2/2)

- Σκοπός είναι να μάθουμε (εκτιμήσουμε) την πραγματική απεικόνιση μεταξύ δεδομένων εισόδου και στόχου, η οποία συμβολίζεται με f^*:

- Ένας αλγόριθμος μάθησης, *A*, είναι μια διαδικασία η οποία με βάση το σύνολο εκπαίδευσης υπολογίζει μια εκτίμηση της συνάρτησης: $\hat{f} = A(\mathscr{D})$

- Η συνάρτηση απώλειας (loss function) ποσοτικοποιεί το κόστος της πρόβλεψης *f(x)* στη θέση του (πραγματικού) *y* και συμβολίζεται με $\ell(y, f(x))$

Στόχος της μάθησης με επίβλεψη

— Δεδομένης μιας συνάρτησης απώλειας (loss function), π.χ. τετραγωνική συνάρτηση, η ποιότητα της εκτίμησης μπορεί να ποσοτικοποιηθεί μέσω του πραγματικού ρίσκου (true risk) (γνωστό και ως σφάλμα γενίκευσης).

$$R(f) = E_{(\mathbf{x},y)\sim p}\left[\left(f(\mathbf{x}) - y\right)^2\right] = \iint_{x\,y} \left(f(\mathbf{x}) - y\right)^2 p(\mathbf{x}, y)\partial y \partial \mathbf{x}\,.$$

— Στόχος της μηχανικής μάθησης με επίβλεψη είναι η επίλυση του παρακάτω προβλήματος βελτιστοποίησης:

$$f^* = \min_{f\in\mathscr{H}} R(f)$$

Στόχος της μάθησης με επίβλεψη

— Ωστόσο, δεν μπορούμε να υπολογίσουμε αναλυτικά το πραγματικό ρίσκο εφόσον η κατανομή p(x,y) από την οποία προέρχονται τα δεδομένα είναι άγνωστη.

Σφάλμα εκπαίδευσης

— Στη πράξη έχουμε πρόσβαση μόνο στο σφάλμα εκπαίδευσης (training error) ή εμπειρικό ρίσκο (empirical risk).

$$\hat{R}(f) = E_{(\mathbf{x},y)\sim D}\left[\left(f(\mathbf{x}) - y\right)^2\right]$$

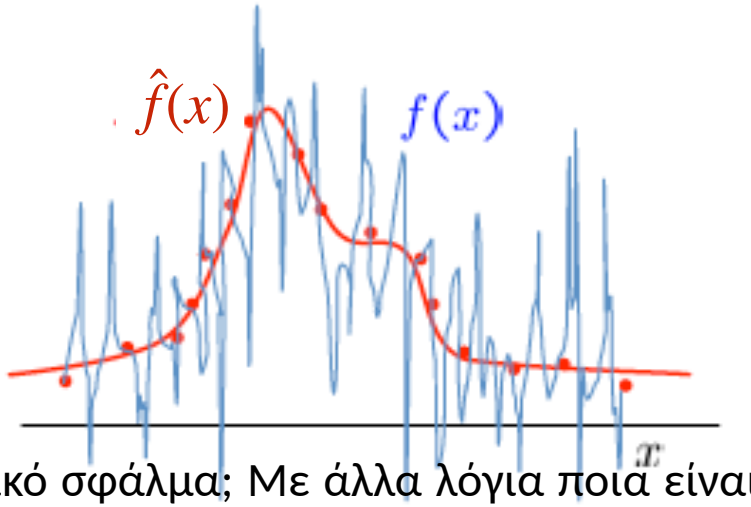
— Αυτό οδηγεί φυσικά στην έννοια της εμπειρικής ελαχιστοποίησης του σφάλματος: Δηλαδή μαθαίνουμε την $\hat{f}$ ελαχιστοποιώντας το $\hat{R}(f)$ ως υποκατάστατο του $R(f)$

$$\hat{f} = \min_{f\in\mathcal{H}} \hat{R}(f)$$

—Ιδανικά θα επιθυμούσαμε: $R(f^*) \approx \hat{R}(\hat{f})$

Q: Ποιες είναι συνέπειες τις ελαχιστοποίησης του εμπειρικού σφάλματος αντί του πραγματικού; Ποια είναι η σχέση μεταξύ αυτών των σφαλμάτων;

Παράδειγμα: Παλινδρόμηση



$$y_i = f(\mathbf{x}_i) + \epsilon_i$$
$$\epsilon \sim \mathcal{N}(0,\sigma^2), i.i.d.$$

- Ποιο είναι το εμπειρικό σφάλμα; Με άλλα λόγια ποια είναι η επίδοση του αλγορίθμου μηχανικής μάθησης στο σύνολο εκπαίδευσης;

  Μηδέν! $\hat{R}(\hat{f}) = 0$

- Τι ισχύει για το πραγματικό ρίσκο (σφάλμα γενίκευσης);

  Μεγαλύτερο του μηδενός: $R(\hat{f}) > 0$

- Το μεγάλο πραγματικό ρίσκο συνεπάγεται την περιορισμένη ικανότητα γενίκευσης του αλγορίθμου μηχανικής στο σύνολο ελέγχου.

Επίδοση αλγορίθμου μηχανικής μάθησης

- Ένας αλγορίθμους μηχανικής μάθησης έχει καλή επίδοση όταν:

    1. Το σφάλμα εκπαίδευσης είναι μικρό.

    2. Η απόσταση μεταξύ σφάλματος εκπαίδευσης και γενίκευσης είναι μικρή. Αυτή η απόσταση είναι γνωστή ώς χασμα γενίκευσης (generalization gap)

- Οι παραπάνω δύο παράγοντες σχετίζονται με δύο εξαιρετικά σημαντικές έννοιες στη μηχανική μάθηση: τις έννοιες underfitting και overfitting.

- Το φαινόμενο underfitting παρατηρείται όταν ο αλγόριθμος μηχανικής μάθησης δεν μπορεί να επιτύχει ικανοποιητικά μικρό σφάλμα εκπαίδευσης.

- Το φαινόμενο overfitting παρατηρείται όταν το χάσμα γενίκευσης είναι πολύ μεγάλο (όπως στο προηγούμενο παράδειγμα).

- Στη πράξη ελέγχουμε τη συμπεριφορά ενός αλγορίθμου ως προς τα φαινόμενα underfitting και overfitting τροποποιόντας τη χωριτικότητα του μοντέλου.

Η χωρητικότητα μοντέλου μάθησης και οι συνέπειες της

- Η χωρητικότητα (πολυπλοκότητα) ενός μοντέλου μηχανικής μάθησης αφορά στην ικανότητα του να μάθει ένα εύρη σύνολο συναρτήσεων μέσω των δεδομένων εκπαίδευσης. Ένας τρόπος για να ελέγξουμε τη χωρητικότητα ενός αλγορίθμου μάθησης είναι με το να τροποποιήσουμε των χώρο υποθέσεων.

- Μοντέλα μικρής χωρητικότητας δεν μπορούν να περιγράψουν σύνθετα δεδομένα ικανοποιητικά (underfitting)

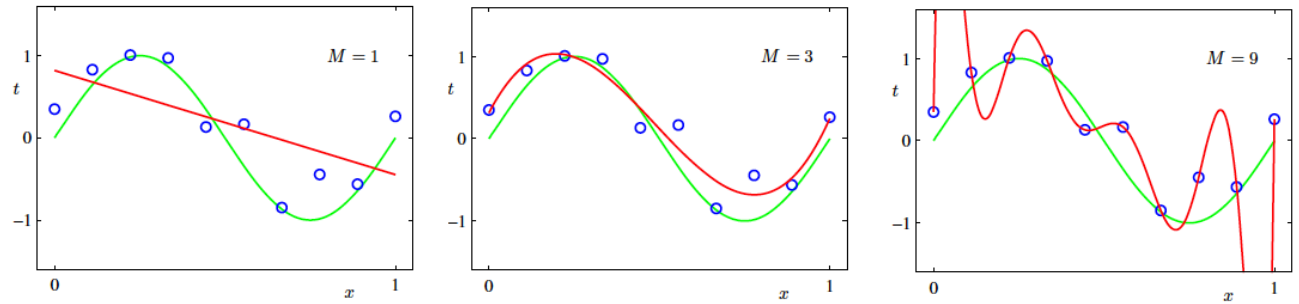
- Μοντέλα μεγάλης χωρητικότητας μπορεί να απομνημονεύσουν ιδιότητες και χαρακτηριστικά του συνόλου εκπαίδευσης τα οποία δεν παρουσιάζονται στο σύνολο ελέγχου (overfitting)

Παραδείγματα χώρων υποθέσεων, under-fitting και overfitting

- Πολυωνυμική παλινδρόμηση με μεταβλητό βαθμό πολυώνυμου, M = 0, 1, 2, … Μεγαλύτερος βαθμός πολυώνυμου => Μεγαλύτερη χωρητικότητα μοντέλου

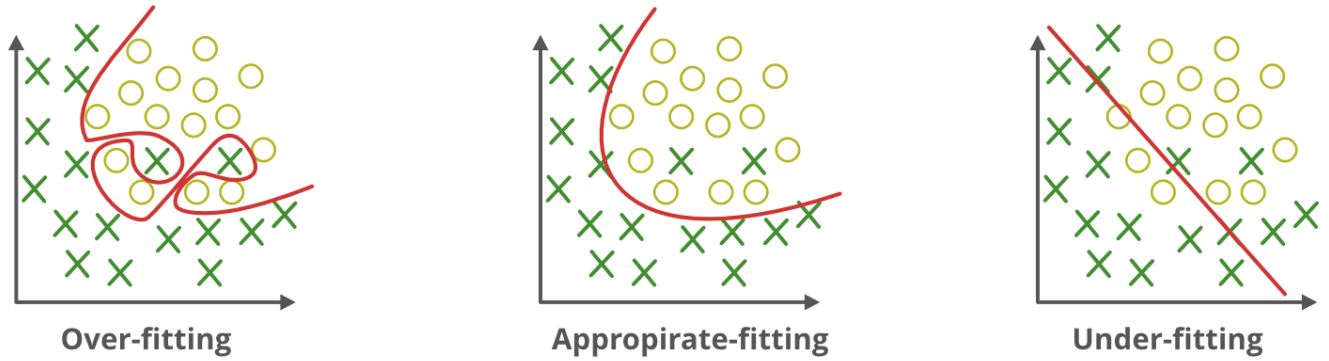
$$\mathcal{H}_{M=1} \prec \mathcal{H}_{M=2} \prec \mathcal{H}_{M=3} \cdots$$

$\mathcal{H}_{M=1}$ $\mathcal{H}_{M=3}$ $\mathcal{H}_{M=9}$

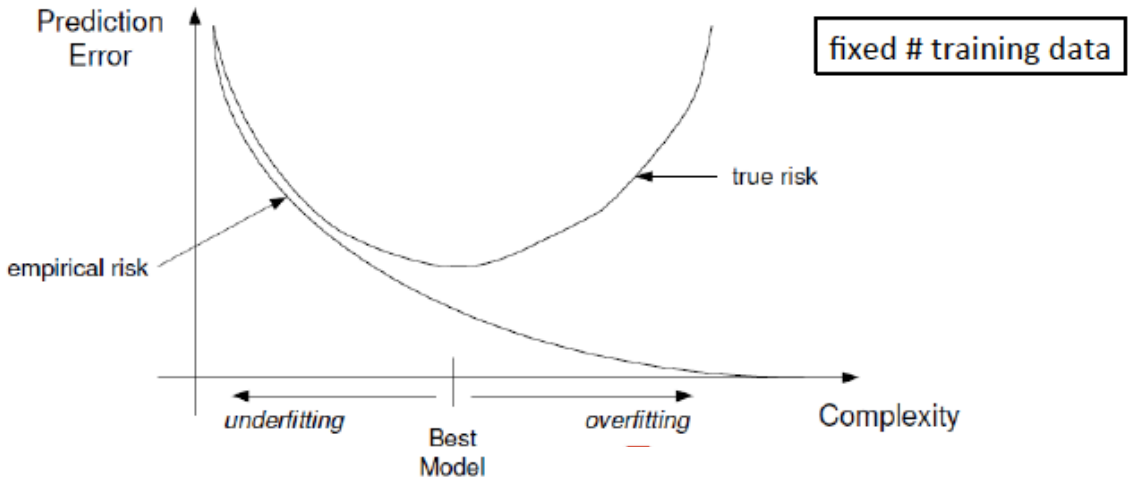


- Ταξινόμηση/Παλινδρόμηση με βάση των κανόνα των K πλησιέστερων γειτόνων K = 1, 2, …
  - Μικρή γειτονιά => Μεγαλύτερη χωρητικότητα μοντέλου

$$\mathcal{H}_{K=1} \succeq \mathcal{H}_{K=2} \succeq \mathcal{H}_{K=3} \cdots$$



Σχέση εμπειρικού και πραγματικού ρίσκου και χωρητικότητας



- Το σφάλμα εκπαίδευσης (empirical risk) είναι μια φθίνουσα συνάρτηση της χωρητικότητας (πολυπλοκότητας) του μοντέλου εφόσον το πλήθος των δεδομένων εκπάιδευσης είναι σταθερό.
- Το σφάλμα ελέγχου ή γενίκευσης (true risk) είναι μιά κυρτή συνάρτηση της χωρητικότητας του μοντέλου.
- Εάν αυξήσουμε τη πολυπλοκότητα του μοντέλου αρκετά το σφάλμα εκπαίδευσης δεν είναι καλός δείκτης για τη συμπεριφορά του πραγματικού ρίσκου.

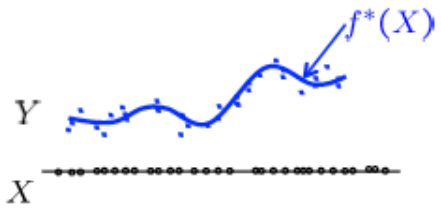
- Συμπέρασμα: Το βέλτιστο ως προς τη πολυπλοκότητα και γενίκευση μοντέλο υπάρχει ωστόσο δεν μπορεί να προσδιοριστεί μελετώντας μόνο το σφάλμα εκπαίδευσης.

Πηγές σφάλματος στο πραγματικό ρίσκο

- Το πραγματικό ρίσκο μπορεί να γραφεί ως άθροισμα τριών συνιστωσών, οι οποίες εκφράζουν διαφορετικές πηγές σφάλματος στην εκτιμώμενη συνάρτηση.

Παράδειγμα: Παλινδρόμηση

$$y_i = f^*(\mathbf{x}_i) + \epsilon_i \quad \epsilon \sim \mathcal{N}(0,\sigma^2), i.i.d.$$

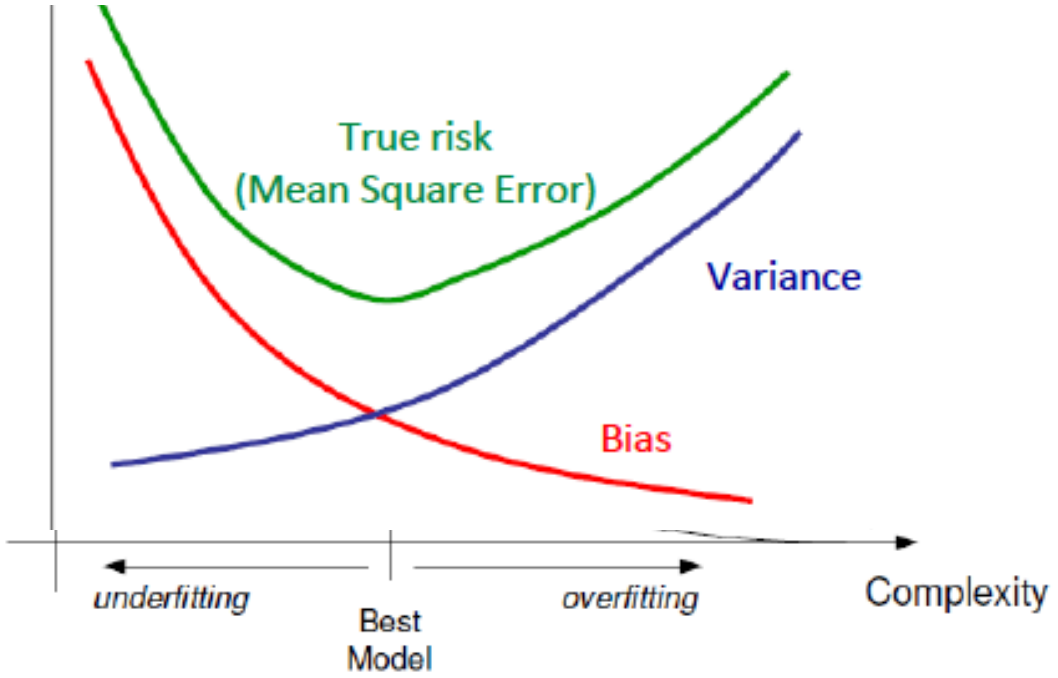


$$R(\hat{f}) = E\left[\left(\hat{f}(\mathbf{x}) - y\right)^2\right] = \underbrace{E\left[\left(E[\hat{f}(\mathbf{x})] - f^*(\mathbf{x})\right)^2\right]}_{\text{Bias}^2} + \underbrace{E\left[\left(\hat{f}(\mathbf{x}) - E[\hat{f}(\mathbf{x})]\right)^2\right]}_{\text{Variance}} + \underbrace{\sigma^2}_{\text{Noise}}$$

- Η μεροληψία (bias) μετράει την αναμενόμενη απόκλιση της εκτίμησης από την πραγματική συνάρτηση. $\text{Bias}(\hat{f}) = E[\hat{f}] - f^*$
- Η διακύμανση (variance) ποσοτικοποιεί την απόκλιση του εκτιμητή από τη μέση τιμή του, η οποία προκαλείται από τη χρήση διαφορετικών συνόλων εκπαίδευσης ή/και μοντέλων διαφορετικής πολυπλοκότητας. $\text{Variance}(\hat{f}) = E[(\hat{f} - E[\hat{f}])^2]$
- Ο παράγοντας θορύβου σχετίζεται με το θόρυβο που εγγενώς υπάρχει στα δεδομένα.

Συμβιβασμός Μεροληψίας - Διακύμανσης

$$R(\hat{f}) = E\left[\left(\hat{f}(\mathbf{x}) - y\right)^2\right] = \underbrace{E\left[\left(E[\hat{f}(\mathbf{x})] - f^*(\mathbf{x})\right)^2\right]}_{Bias^2} + \underbrace{E\left[\left(\hat{f}(\mathbf{x}) - E[\hat{f}(\mathbf{x})]\right)^2\right]}_{Variance} + \underbrace{\sigma^2}_{\text{Noise}}$$



Υπερπαράμετροι μοντέλων μάθησης

- Οι περισσότεροι αλγόριθμοι μηχανικής μάθησης έχουν υπερ-παραμέτρους (hyperparameters), δηλαδή παραμέτρους οι οποίες προσδιορίζουν την συμπεριφορά ενός αλγορίθμου.

Για παράδειγμα: ο βαθμός του πολυώνυμου μίας παραμετρικής πολυωνυμικής συνάρτησης δρα ως υπερπαράμετρος χωρητικότητας, ελέγχοντας πρακτικά τη πολυπλοκότητα του χώρου υποθέσεων του μοντέλου μάθησης. Το πλήθος των γειτόνων στη μη παραμετρική μέθοδο ταξινόμησης/παλινδρόμησης των πλησιέστερων γειτόνων λειτουργεί ως παράμετρος χωρητικότητας.

Ε: Πως επιλέγουμε τις υπερπαραμέτρους ώστε να το σφάλμα γενίκευσης (πραγματικό ρίσκο) να είναι ελάχιστο;

Ε: Πως επιλέγουμε το βέλτιστο μοντέλο ως προς τη χωρητικότητα του οποίου η επίδοση είναι καλή σε δεδομένα εκτός του συνόλου εκπαίδευσης;

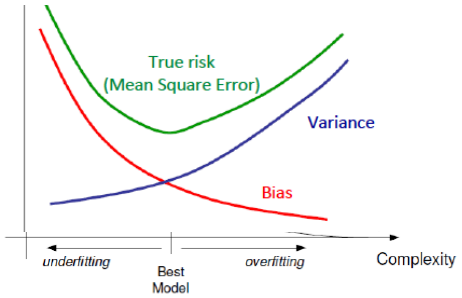
Επιλογή μοντέλου

- Έστω ένα σύνολο μοντέλων μάθησης διαφορετικής πολυπλοκότητας (χωρητικότητας) τα οποία διατάσσονται ως προς αύξουσα πολυπλοκότητα.

$$\{\mathcal{H}_\lambda\}_{\lambda\in\Lambda}, \mathcal{H}_1 \prec \mathcal{H}_2 \prec \cdots$$

- Το επιθυμητό μοντέλο είναι αυτό έχει την μικρότερη πολυπλοκότητα και ελαχιστοποιεί το σφάλμα γενίκευσης (πραγματικό ρίσκο)

$$\hat{f} = \min_{\lambda} \min_{f\in\mathcal{H}_\lambda} R(f)$$



- Πρακτικά, μπορούμε να επιλέξουμε το μοντέλο καταλλήλης πολυπλοκότητας χωρίζοντας τα διαθέσημα δεδομένα σε υποσύνολα εκπαίδευσης και επικύρωσης (validation set). Τα σύνολο επικύρωσης χρησιμοποιούνται για να εκτιμήσουμε την τιμή του πραγματικού ρίσκου.

Μέθοδος Hold-out

- Στόχος: να βρούμε το μοντέλο με το μικρότερο σφάλμα γενίκευσης
- Σύνολο δεδομένων: $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N} \sim p$
- Βήμα 1: Χωρίζουμε τα διαθέσιμα δεδομένα σε δύο σύνολα:
  - Σύνολο εκπαίδευσης: $\mathscr{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{M}$
  - Σύνολο επικύρωσης: $\mathscr{D}_V = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=M+1}^{N}$
- Βήμα 2: Χρησιμοποιούμε το σύνολο εκπαίδευσης για να εκπαιδεύσουμε ένα μοντέλο για κάθε κλάση πολυπλοκότητας (υποθέσεων)

$$\hat{f}_\lambda = \arg\min_{f \in \mathscr{H}_\lambda} \hat{R}(f) \rightarrow \{\hat{f}_{\lambda=1}, \hat{f}_{\lambda=2}, \cdots\}$$

- Βήμα 3: Χρησιμοποιούμε το σύνολο επικύρωσης για να διαλέξουμε το μοντέλο με το μικρότερο σφάλμα γενίκευσης

$$\hat{\lambda} = \arg\min_{\lambda \in \Lambda} \hat{R}_V(\hat{f}_\lambda)$$

- Το βέλτιστο μοντέλο που παράγει αυτή η μέθοδος είναι: $\hat{f} = \hat{f}_{\hat{\lambda}}$

Μειονεκτήματα: 1) Το πλήθος των διαθέσιμων δεδομένων μπορεί να είναι μικρό 2) Το σύνολο επικύρωσης μπορεί να μην είναι αντιπροσωπευτικό του συνόλου ελέγχου.

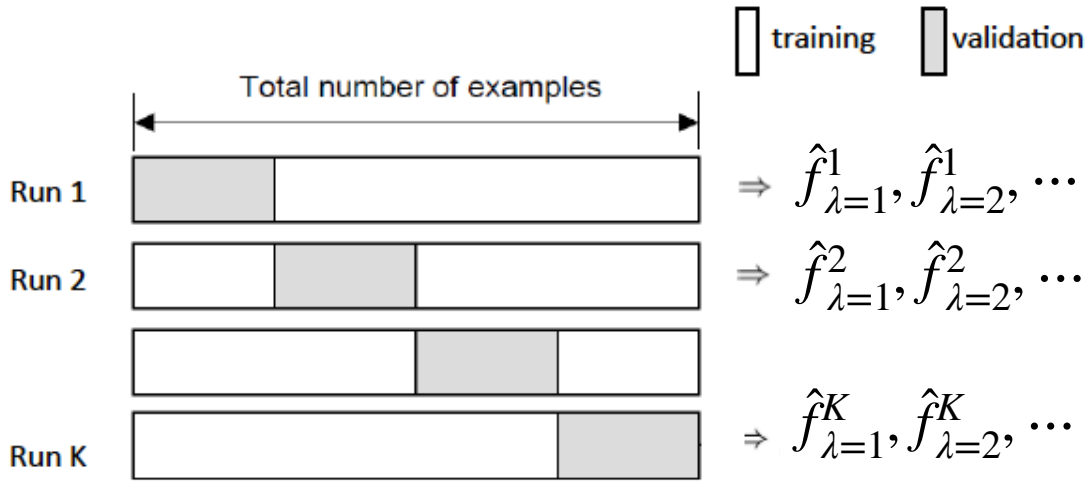
Μέθοδος Cross-Validation

- Τα μειονεκτήματα της μεθόδου Hold-out για την επιλογή μοντέλου μπορούν να αντιμετωπιστούν χρησιμοποιώντας μεθόδους τυχαίας υποδειγματοληψίας των δεδομένων, αυξάνοντας όμως το υπολογιστικό κόστος της όλης διαδικασίας.

- Μέθοδος K-Fold Cross-Validation

- Βήμα 1: Χωρίζουμε τα διαθέσιμα δεδομένα σε K σύνολα (folds):

- Βήμα 2: Για κάθε κλάση πολυπλοκότητας λ εκπαιδεύουμε K εκτιμήσεις χρησιμοποιώντας τα K-1 σύνολα για εκπαίδευση και 1 σύνολο για επικύρωση. Ουσιαστικά εκτελούμε το βήμα 2 της μεθόδου Hold out K φορες.



Μέθοδος Cross-Validation

- Βήμα 3: Χρησιμοποιούμε τα K σύνολα επικύρωσης για να διαλέξουμε το μοντέλο με το μικρότερο κατά μέσο όρο σφάλμα γενίκευσης.

$$\hat{\lambda} = \arg\min_{\lambda \in \Lambda} \frac{1}{K} \sum_{k=1}^{K} \hat{R}_{V_k}(\hat{f}_{\lambda}^{k})$$

- Η μέθοδος επιστρέφει το μοντέλο που κατά μέσο όρο στα K πειράματα έχει μικρότερο σφάλμα γενίκευσης.

- Συνήθως διαλέγουμε K = 10