Μηχανική Μάθηση

Μιχάλης Τίτσιας

Διάλεξη 2ή

Η τεχνική *cross validation*, εισαγωγή στην θεωρία πιθανοτήτων και στα πιθανοτικά μοντέλα

Περιεχόμενα

- Σύντομη επανάληψη από τα προηγούμενα
- Cross-validation
- Αβεβαιότητα στα προβλήματα μηχανικής μάθησης
- Επανάληψη στην θεωρία πιθανοτήτων
- Πιθανοτικό μοντέλο για παλινδρόμηση
- Μέθοδος μέγιστης πιθανοφάνειας (maximum likelihood)

Η γενική δομή ενός συστήματος μηχανικής μάθησης

Ένα σύστημα μηχανική μάθησης αποτελείται από

Δεδομένα:

- Συλλογή και προεπεξεργασία δεδομένων (feature selection/extraction)

Μοντέλο ή υπόθεση:

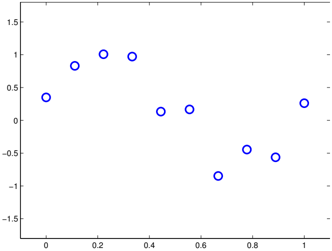
- Π.χ. η γραμμική ή η τετραγωνική συνάρτηση στο πρόβλημα παλινδρόμησης
- Εξαρτάται από άγνωστους παραμέτρους

Αλγόριθμοι εκπαίδευσης:

- Συναρτήσεις κόστους βάσει των οποίων μαθαίνουμε τις άγνωστες παραμέτρους του μοντέλου
- Αλγόριθμοι βελτιστοποίησης

Η γενική δομή ενός συστήματος μηχανικής μάθησης

Παράδειγμα



Δεδομένα: Φαίνονται στο σχήμα

Μοντέλο ή υπόθεση:

$$y(x, \mathbf{w}) = w_0 + w_1 x + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Αλγόριθμος εκπαίδευσης:

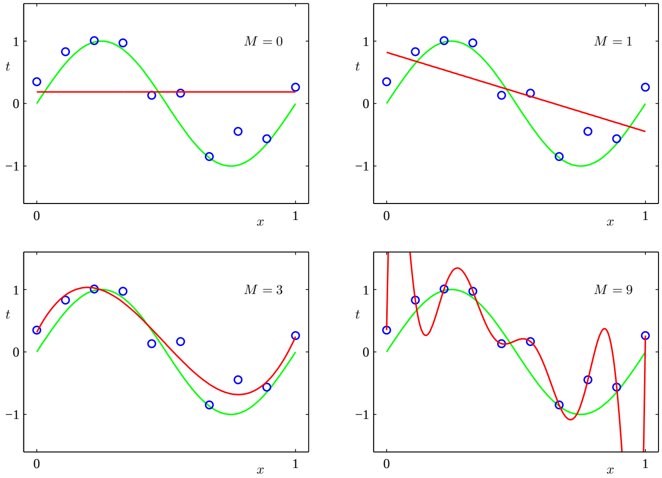
- Συναρτήση κόστους $\Rightarrow E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \left(y(x_n, \mathbf{w}) - t_n\right)^2$
- Αλγόριθμος βελτιστοποίησης $\Rightarrow$ λύση ενός γραμμικού συστήματος (θα το δούμε στο επόμενο μάθημα)

Υπερεκπαίδευση, υποεκπαίδευση

Ένα σημαντικό θέμα αφορά την επιλογή μοντέλου (model selection) ώστε να επιτυγχάνουμε την

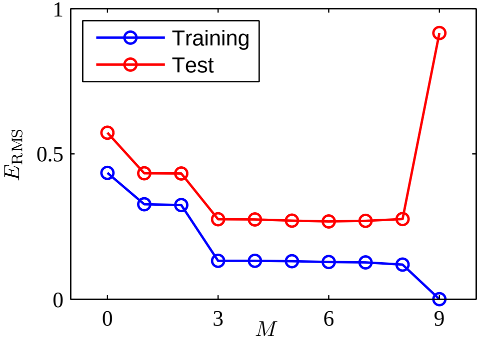
- αποφυγή των φαινομένων υπερεκπαίδευσης (overfitting) και υποεκπαίδευσης (underfitting)

Υπερεκπαίδευση, υποεκπαίδευση



Το $M = 9$ μοντέλο είναι υπερεκπαιδευμένο (overfitted)
Τα $M = 0, 1$ μοντέλα είναι υποεκπαιδευμένα (underfitted)
Το $M = 3$ μοντέλο είναι το καλύτερο

Υπερεκπαίδευση, υποεκπαίδευση



Επίδοση στα δεδομένα ελέγχου: Μέσο-σφάλμα (root-mean-square-error)

$$\sqrt{\frac{\sum_{x_*}(y(x_*, \mathbf{w}^*) - t_*)^2}{\text{Αριθμός δεδομένων ελέγχου}}}$$

(τα υποεκπαιδευμένα και υπερεκπαιδευμένα μοντέλα δεν έχουν καλή επίδοση)

Κανονικοποίηση (regularization)

- Μια έξυπνη πολιτική στην κατασκευή συστημάτων μηχανικής μάθησης
- είναι η χρήση **πολύ ευέλικτων μοντέλων** (ως default!)
    - $\Rightarrow$ που ενδεχομένως θα μπορούσαν να επιλύσουν και τα πιο σύνθετα προβλήματα
- Έπειτα θα θέλαμε κατά περίπτωση να **προσαρμόζουμε/περιορίζουμε την ευελιξία** των μοντέλων αυτών
    - $\Rightarrow$ ώστε να αποφεύγεται η υπερεκπαίδευση

Κανονικοποίηση

Κανονικοποίηση (regularization) των παραμέτρων $\mathbf{w}$

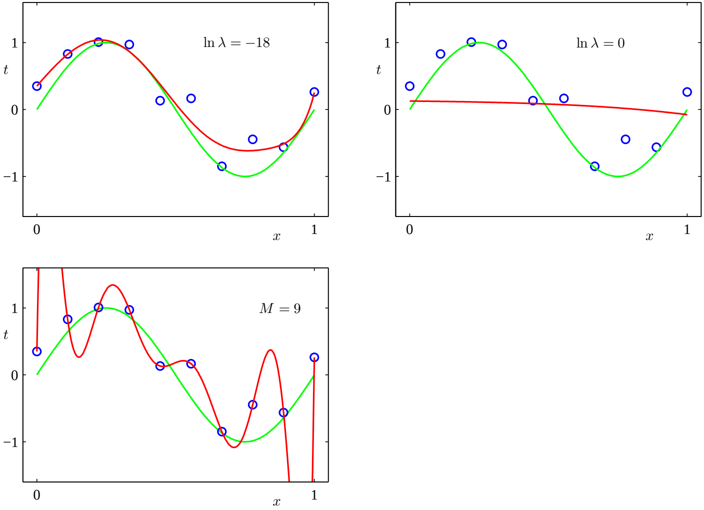
- Θα θέλαμε μια νέα συνάρτηση κόστους που να αποτρέπει μεγάλες τιμές των παραμέτρων

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(y(x_n, \mathbf{w}) - t_n)^2 + \lambda\frac{||\mathbf{w}||^2}{2}$$

όπου $||\mathbf{w}||^2 = w_0^2 + w_1^2 + \ldots + w_M^2$ και $\lambda > 0$

- Ο όρος $\lambda\frac{||\mathbf{w}||^2}{2}$ ‘τιμωρεί’ μεγάλες τιμές των παραμέτρων
- λ ονομάζεται παράμετρος κανονικοποίησης

Κανονικοποίηση



Το $M = 9$ μοντέλο εκπαιδευμένο για διαφορετικές τιμές του λ. Για κάποια τιμή του λ το μοντέλο φαίνεται ιδανικό!

Κανονικοποίηση

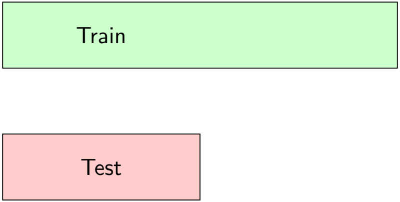
$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(y(x_n,\mathbf{w}) - t_n)^2 + \lambda\frac{||\mathbf{w}||^2}{2}$$

Ερωτήσεις:

1. Πώς επιλέγουμε την τιμή του λ;
2. Ποια είναι η ερμηνεία πίσω από την χρήση της $E(\mathbf{w})$; (θα μπορούσε η $E(\mathbf{w})$ να είχε άλλη μορφή;)

Θα ξεκινήσουμε με το ερώτημα 1) και θα παρουσιάσουμε μια τεχνική που μας επιτρέπει να βρίσκουμε κατάλληλες τιμές για το λ;

Cross-validation



$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(y(x_n, \mathbf{w}) - t_n)^2 + \lambda\frac{||\mathbf{w}||^2}{2}$$

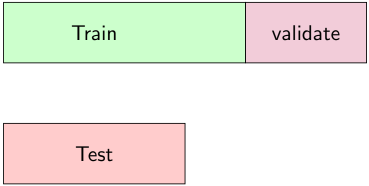
- Η μοντέλο εκπαιδεύεται χρησιμοποιώντας τα δεδομένα εκπαίδευσης
- Το πόσο καλό είναι το μοντέλο εξαρτάται από τα δεδομένα ελέγχου τα οποία είναι άγνωστα κατά την εκπαίδευση

Ιδανικά θα θέλουμε να επιλέξουμε εκείνο το λ για οποίο επιτυγχάνουμε την καλύτερη δυνατη πρόβλεψη στα δεδομένα ελέγχου

- Ωστόσο τα δεδομένα ελέγχου δεν τα γνωρίζουμε

Cross-validation

Simple validation

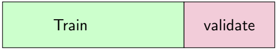


Ιδέα: Κατασκεύασε τεχνητά ένα σύνολο ελέγχου

- Χωρίσε το σύνολο εκπαίδευσης σε δύο κομμάτια: σύνολο εκπαίδευσης και **σύνολο αξιολόγησης**
- Εκπαίδευσε το μοντέλο μόνο με το πρώτο κομμάτι
- Μέτρα την επίδοση στο **σύνολο αξιολόγησης**
- Επέλεξε εκείνη την τιμή του λ (ή ο,τιδήποτε άλλο καθορίζει την πολυπλοκότητα του μοντέλου π.χ. τάξη του πολυωνύμου M) που οδηγεί στην καλύτερη επίδοση στο σύνολο αξιολόγησης

Cross-validation

Simple validation



Συχνά χρησιμοποιούμε 80% από τα δεδομένα για εκπαίδευση και 20% για έλεγχο. Στο παράδειγμα μας η τεχνική ακολουθεί τα βήματα

1. Χώρισε τα δεδομένα σε 80% για εκπαίδευση (σύνολο T) και 20% για έλεγχο (V)
2. Έστω ένα σύνολο από λς: $\{\lambda_1, \lambda_2, \ldots\}$
3. Για λ_i εκτέλεσε τα βήματα 4 και 5
4. Χρησιμοποιώντας το σύνολο εκπαίδευσης βρες $\mathbf{w}_*$ που ελαχίστοποιεί $E(\mathbf{w}) = \frac{1}{2}\sum_{n\in T}^{N}(y(x_n, \mathbf{w}) - t_n)^2 + \lambda_i \frac{||\mathbf{w}||^2}{2}$
5. Μέτρα επίδοση $E_i = \sqrt{\frac{\sum_{n\in V}(y(x_n,\mathbf{w}_*)-t_n)^2}{|V|}}$
6. Επέλεξε λ_{i_*} για το οποίο E_{i_*} είναι το μικρότερο
7. Για το λ_{i_*} που επιλέχθηκε επανέλαβε την εκπαίδευση χρησιμοποιώντας όλα τα δεδομένα (δηλ. την ένωση του T και V)

Cross-validation

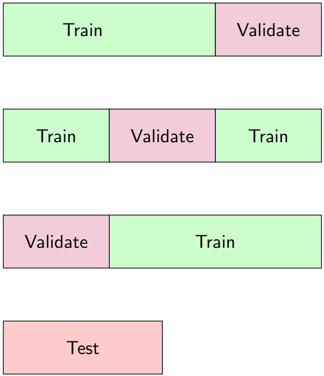
Simple validation

Train	validate

1. Χώρισε τα δεδομένα σε 80% για εκπαίδευση (σύνολο T) και 20% για έλεγχο (V)
2. Έστω ένα σύνολο από λς: $\{\lambda_1, \lambda_2, \ldots\}$
3. Για λ_i εκτέλεσε τα βήματα 4 και 5
4. Χρησιμοποιώντας το σύνολο εκπαίδευσης βρες $\mathbf{w}_*$ που ελαχιστοποιεί $E(\mathbf{w}) = \frac{1}{2}\sum_{n\in T}^{N} \left(y(x_n, \mathbf{w}) - t_n\right)^2 + \lambda_i \frac{||\mathbf{w}||^2}{2}$
5. Μέτρα επίδοση $E_i = \sqrt{\frac{\sum_{n\in V}(y(x_n,\mathbf{w}_*)-t_n)^2}{|V|}}$
6. Επέλεξε λ_{i_*} για το οποίο E_{i_*} είναι το μικρότερο
7. Για το λ_{i_*} που επιλέχθηκε επανέλαβε την εκπαίδευση χρησιμοποιώντας όλα τα δεδομένα (δηλ. την ένωση του T και V)

Μειονέκτημα: Τα δεδομένα μπορεί να είναι πολύ λίγα

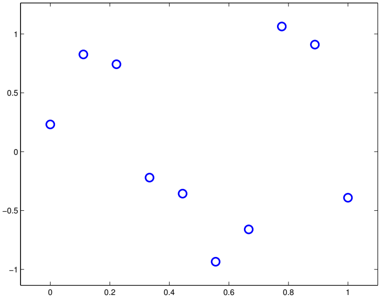
Cross-validation



- Το σύνολο εκπαίδευσης χωρίζεται σε K-κομμάτια: τρέχουμε K φορές τον αλγόριθμο εκπαίδευσης για την ίδια τιμή του λ_i (ή οποιαδήποτε άλλη παράμετρο πολυπλόκοτητας μοντέλου) χρησιμοποιώντας $K-1$ κομμάτια για εκπαίδευση και το κομμάτι που απομένει για αξιολόγηση
- Η μέση τιμή επίδοσης από τα K τρεξίματα χρησιμοποιείται για την αξιολόγηση τους μοντέλου για το συγκεκριμένο λ_i

Cross-validation

Παράδειγμα

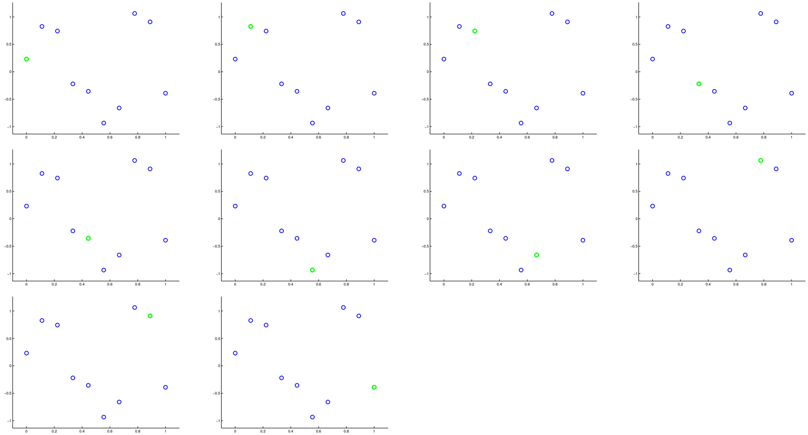


$$y(x, \mathbf{w}) = w_0 + w_1 x + \ldots + w_9 x^9 = \sum_{j=0}^{9} w_j x^j$$

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} (y(x_n, \mathbf{w}) - t_n)^2 + \lambda \frac{||\mathbf{w}||^2}{2}$$

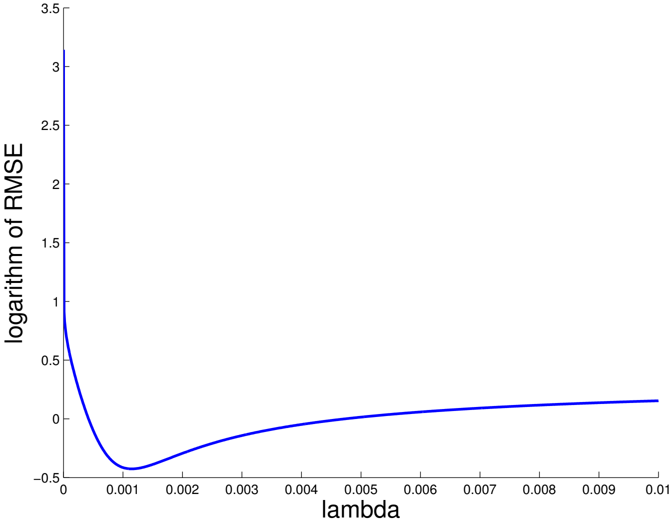
Θα θέλαμε να εφαρμόσουμε cross-validation για να επιλέξουμε το λ

Παράδειγμα



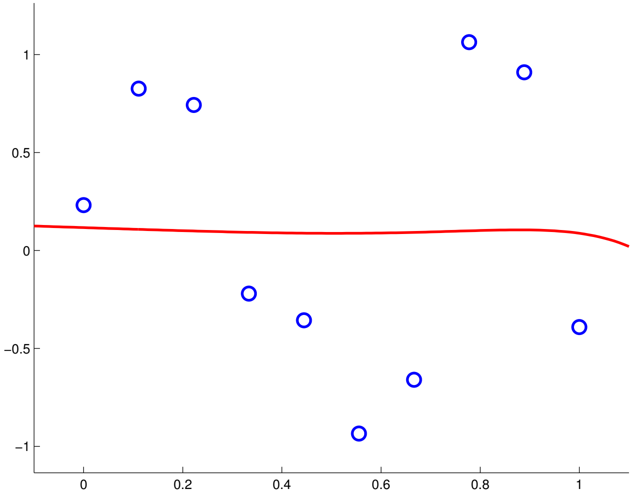
Χωρίζουμε τα 10 δεδομένα σε 10 κομμάτια. Αυτή η ειδική περίπτωση του cross-validation ονομάζεται leave-one-out. Θα εξετάσουμε διάφορες τιμές του λ από την τιμή 0 ως την τιμή 1

Cross-validation



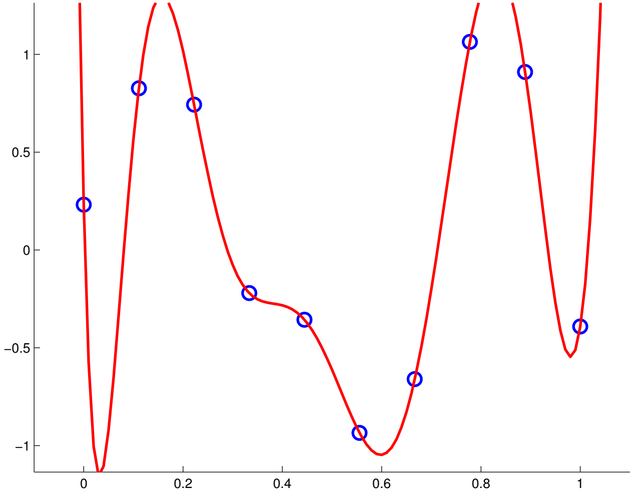
Επίδοση $E_i = \frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{\sum_{n\in V_k}(y(x_n,\mathbf{w}_*)-t_n)^2}{|V_k|}} = \frac{1}{N}\sum_{n=1}^{N}|y(x_n,\mathbf{w}_*)-t_n|$
(αφού το V_k περιέχει μόνο ένα δεδομένο) συναρτήσει των τιμών λ_i

Cross-validation



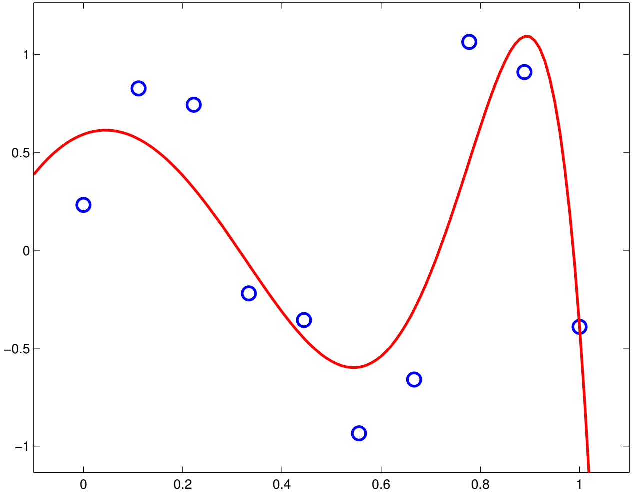
Πολύ μεγάλο λ

Cross-validation



$\lambda = 0$

Cross-validation



Το μοντέλο με το βέλτιστο $\lambda = 0.0012$

Cross-validation

Πλεονεκτήματα

- Cross-validation είναι μια γενική μέθοδος για αποφυγή του overfitting και αξιολόγηση μοντέλων
- Όταν αναζητούμε μια παράμετρο κανονικοποίησης, η εφαρμογή της μεθόδου είναι αρκετά γρήγορη

Μειονέκτημα

- Όταν έχουμε πολλούς παραμέτρους κανονικοποίησης, η μέθοδος γίνεται υπερβολικά δαπανηρή (ουσιαστικά μη εφαρμόσιμη)

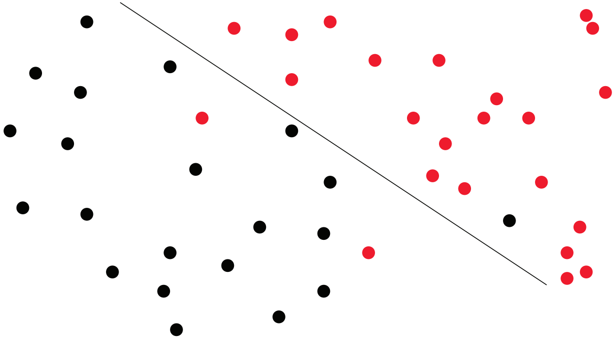
Αβεβαιότητα στα προβλήματα μηχανικής μάθησης

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(y(x_n, \mathbf{w}) - t_n)^2 + \lambda\frac{||\mathbf{w}||^2}{2}$$

- Ποια είναι η ερμηνεία πίσω από την χρήση της $E(\mathbf{w})$; και συγκεκριμένα του όρου $\frac{1}{2}\sum_{n=1}^{N}(y(x_n, \mathbf{w}) - t_n)^2$;
- Η συγκεριμένη μορφή της $E(\mathbf{w})$ μήπως υποννοεί κάποιου είδους υπόθεσης για τη στοχαστικότητας ή αβεβαιότητας ή θορύβου που χαρακτηρίζει τα δεδομένα

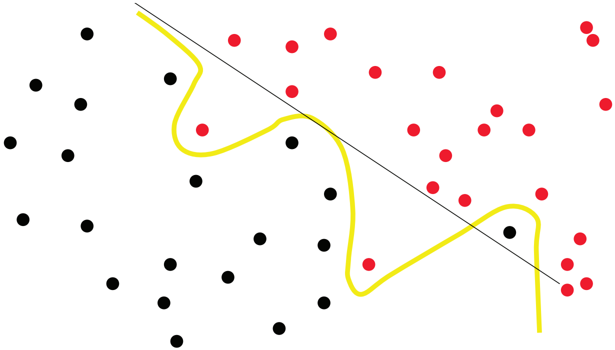
Η αβεβαιότητα είναι ένα γενικό χαρακτηριστικό των προβλημάτων μηχανικής μάθησης. Υπάρχουν πολλαπλοί παράμετροι που έχουν ως συνέπεια την αβεβαιότητα

Αβεβαιότητα στα προβλήματα μηχανικής μάθησης



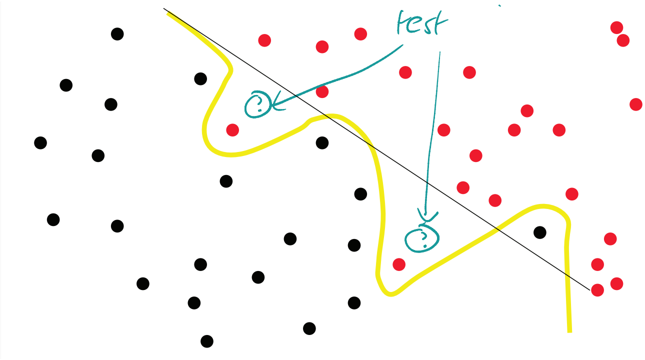
Ποιο μοντέλο είναι καλύτερο;

Αβεβαιότητα στα προβλήματα μηχανικής μάθησης



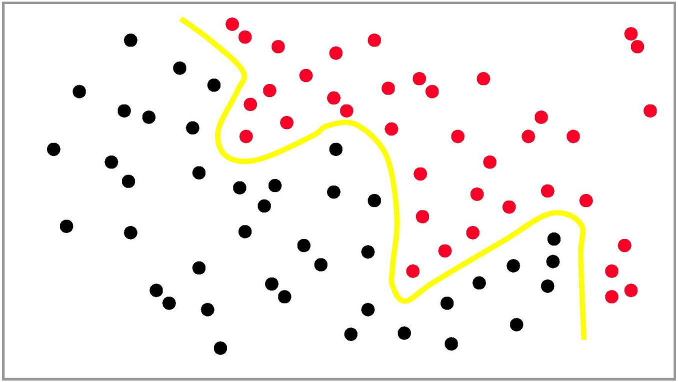
Ποιο μοντέλο είναι καλύτερο; Η **μαύρη** ή η κίτρινη γραμμή;

Αβεβαιότητα στα προβλήματα μηχανικής μάθησης



Η επίδοση σε άγνωστα δεδομένα είναι αυτή που μετράει

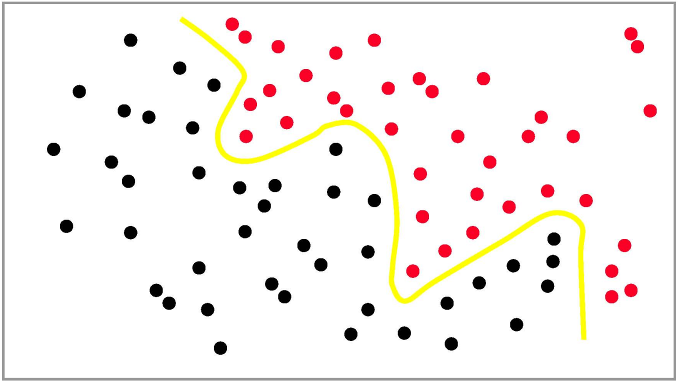
Αβεβαιότητα στα προβλήματα μηχανικής μάθησης



Αν είχαμε μεγαλύτερο δείγμα δεδομένων ενδεχομένως να είμασταν πιο σύγουροι για το ποιο μοντέλο είναι καλύτερο

- Το ότι έχουμε ένα συγκεκριμένο και πεπερασμένο δείγμα αποτελεί πηγή αβεβαιότητας
- Αν είχαμε παρά πόλλα (άπειρα) δεδομένα, τότε θα είχαμε πλήρη πληροφορία για το πρόβλημα

Αβεβαιότητα στα προβλήματα μηχανικής μάθησης



- Κατά κάποιο τρόπο το πρόβλημα μας είναι ένα πρόβλημα στατιστικής ανάλυσης
- Δηλ. από το δείγμα δεδομένων θα θέλαμε να βρούμε κατάλληλα μοντέλα που γενικεύουν καλά σε όλο τον πληθυσμό από τον οποίο το δείγμα έχει προέρθει

Αβεβαιότητα στα προβλήματα μηχανικής μάθησης

Πηγές της αβεβαιότητας:

- Θόρυβος στα δεδομένα
- Το ότι έχουμε ένα συγκεκριμένο δείγμα δεδομένων
- Μερική ή καθόλου γνώση για το ποια μέθοδος/μοντέλο επίλυσης του προβλήματος είναι κατάλληλη

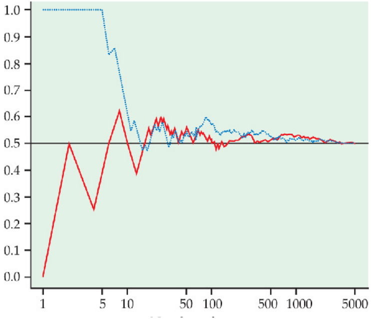
Η επίστημη της αβεβαιότητας είναι η θεωρία πιθανότητων

- η οποία αποτελεί το θεωρητικό υπόβαθρο κατασκευής συστηματών μηχανικής μάθησης

Επανάληψη στην θεωρία πιθανοτήτων

- Έστω ότι ρίχνουμε ένα νόμισμα ή ένα ζάρι μια φορά
  - Δεν μπορούμε να προβλέψουμε το αποτέλεσμα
  - οπότε το αποτέλεσμα είναι τυχαίο
- Ωστόσο αν ρίξουμε το νόμισμα πολλές φορές εμφανίζεται μια κανονικότητα
  - Που μπορεί να οδηγήσει σε βέβαιη πρόβλεψη κάποιων πραγμάτων
- Αυτή η κανονικότητα που εμφανίζεται όταν επαναλάβουμε το πείραμα πολλές φορες είναι η ιδέα πίσω από τη θεωρία πιθανοτήτων

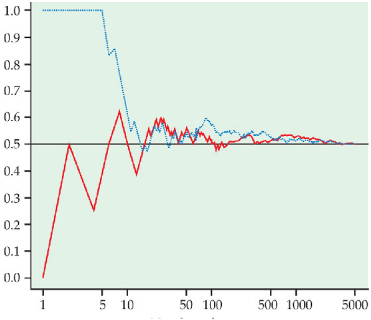
Επανάληψη στην θεωρία πιθανοτήτων



Σχήμα: Οριζόντιος άξονας αντιστοιχεί στο αριθμό των φορών που ρίχτηκε ένα δίκαιο νόμισμα και ο κάθετος άξονας στο ποσοστό (από 0 ώς 1) που το αποτέλεσμα ήταν κορώνα.

- Έστω ότι επαναλαμβάνουμε δύο φορές το ακόλουθο πείραμα: Ρίψη 5000 φορών ενός νομίσματος
- Η κόκκινη γραμμή αντιστοιχεί στη πρώτη επανάληψη του πειράματος και η μπλέ γραμμή στη δεύτερη επανάληψη

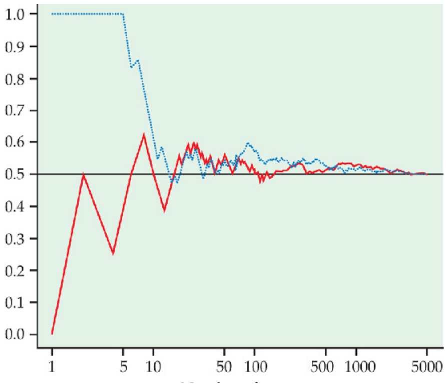
Επανάληψη στην θεωρία πιθανοτήτων



Σχήμα: Οριζόντιος άξονας αντιστοιχεί στο αριθμό των φορών που ρίχτηκε ένα δίκαιο νόμισμα και ο κάθετος άξονας στο ποσοστό (από 0 ώς 1) που το αποτέλεσμα ήταν κορώνα.

- Παρατηρούμε ότι αρχικά οι δύο γραμμές έχουν διαφορετική συμπεριφορά (π.χ. η μπλέ είναι ίση με 1 για τις 5 πρώτες φορές που σημαίνει ότι οι 5 πρώτες ρίψεις ήταν κορώνα)
- Για μεγάλο αριθμό ρίψεων οι δύο γραμμές τείνουν στο 0.5

Επανάληψη στην θεωρία πιθανοτήτων



- Τελικά μπορούμε να πούμε με βεβαιότητα ότι ο λόγος ή ποσοστό των φορών που έρχεται κορώνα είναι 0.5
- Δηλαδή μια κανονικότητα διαφαίνεται στην επανάληψη των πολλών φορών

Τυχαίο πείραμα: Ένα πείραμα ή φαινόμενο είναι τυχαίο όταν δεν μπορούμε να προβλέψουμε ακριβώς το αποτέλεσμα. Ωστόσο υπάρχει μια κανονικότητα που διαφαίνεται όταν επαναλάβουμε το πείραμα πολλές φορές.

Πιθανότητα: του κάθε αποτελέσματος του τυχαίου πειράματος είναι το ποσοστό ή αναλογία (εκφρασμένη στο διάστημα 0 έως 1) των φορών που το αποτέλεσμα θα συμβεί σε μια μεγάλη σειρά επαναλήψεων του πειράματος

Επανάληψη στην θεωρία πιθανοτήτων

Διαισθητικός ορισμός της πιθανότητας

$$\text{Πιθανότητα του } j = \lim_{N \to \infty} \frac{n_j}{N}$$

όπου N ο αριθμός των επαναλήψεων του πειράματος και n_j ο αριθμός των φορών που το αποτέλεσμα ήταν j

- Σε ορισμένες περιπτώσεις μπορούμε να κατανοήσουμε διαισθητικά την πιθανοτήτα μέσω «συμμετρίας»
  - Ένα νόμισμα είναι απόλυτα συμμετρικό οπότε η πιθανότητα να έρθει κορώνα είναι 0.5

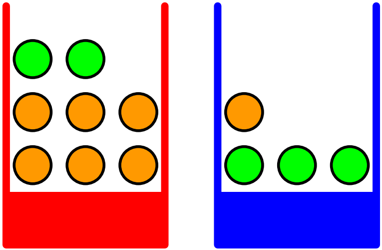
Επανάληψη στην θεωρία πιθανοτήτων

Ορολογία

- Τυχαία μεταβλητή (random variable): Μια μεταβλητή που η τιμή της καθορίζεται μέσω τυχαίου πειράματος
  - Διακριτή τυχαία μεταβλητή: Παίρνει διακριτές τιμές π.χ. $\{0, 1, 2, \ldots\}$
  - Συνεχής τυχαία μεταβλητή: Παίρνει συνεχείς τιμές στο $\mathbb{R}$
- Δειγματικός χώρος (sample space): Το σύνολο τιμών που παίρνει μια τυχαία μεταβλητή
- Ενδεχόμενο (event): Ένα υποσύνολο του δειγματικού χώρου
- Συμπερασματολογία (inference): Εξαγωγή συμπεράσματος για τιμές τυχαίων μεταβλήτων δοθέντος των παρατηρούμενων δεδομένων

Επανάληψη στην θεωρία πιθανοτήτων

Πορτοκάλια και μήλα

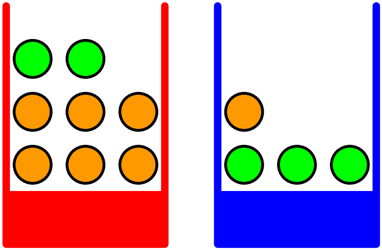


Πείραμα (Παραγωγή ενός δεδομένου):

1. Επέλεξε ένα από δύο κουτιά ώστε το μπλε επιλέγεται με πιθανότητα $\frac{6}{10} = 0.6$
2. Από το κουτί επιλέχθηκε στο 1), επέλεξε ένα φρούτο
   - Αυτό είναι το δεδομένο σου!
3. Το επιλεγμένο φρούτο επιστρέφεται στο κουτί

Πρόβλημα συμπερασματολογίας: Αν επιλέξαμε ένα πορτοκάλι, τότε ποιο ήταν το κουτί από το οποίο προήρθε;

Πορτοκάλια και μήλα

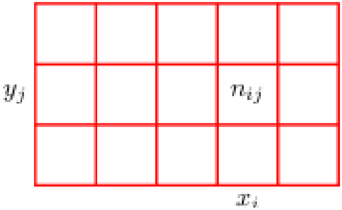


Δύο τυχαίες μεταβλητές:

- X: ταυτοποιεί το κουτί που επιλέχθηκε παίρνωντας τιμές στο $\{x_1, x_2\} = \{red, blue\}$
- Y: καθορίζει το φρούτο και παίρνει τιμές $\{y_1, y_2\} = \{orange, apple\}$
- Για να λύσουμε το πρόβλημα συμπερασματολογίας θα πρέπει να ορίσουμε την από κοινού πιθανότητα

$$P(X = x_i, Y = y_j), \quad i, j = 1, 2$$

Επανάληψη στην θεωρία πιθανοτήτων



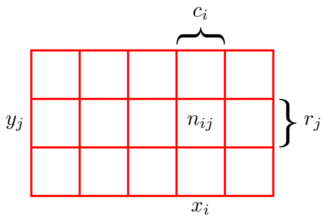
Από κοινού πιθανότητα:

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

n_{ij}: ο αριθμός των φόρων που συγχρόνως η πρώτη μεταβλητή X παίρνει την τιμή x_i και η δεύτερη μεταβλητή Y παίρνει την τιμή y_j

N: Συνολικός αριθμός επανάληψεων του πειράματος (και $N \to \infty$)

Επανάληψη στην θεωρία πιθανοτήτων



- Από κοινού πιθανότητα:

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

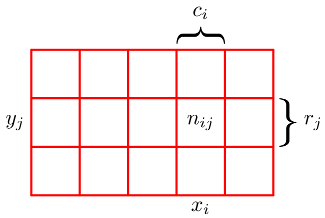
- Περιθωριοποιημένη (marginal) πιθανότητα:

$$P(X = x_i) = \frac{c_i}{N}$$

- Δεσμευμένη ή υπο συνθήκη (conditional) πιθανότητα:

$$P(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

Επανάληψη στην θεωρία πιθανοτήτων



- Ο κανόνας αθροίσματος

$$P(X = x_i) = \frac{c_i}{N} = \frac{\sum_j n_{ij}}{N} = \sum_j P(X = x_i, Y = y_j)$$

- Κανόνας γινομένου

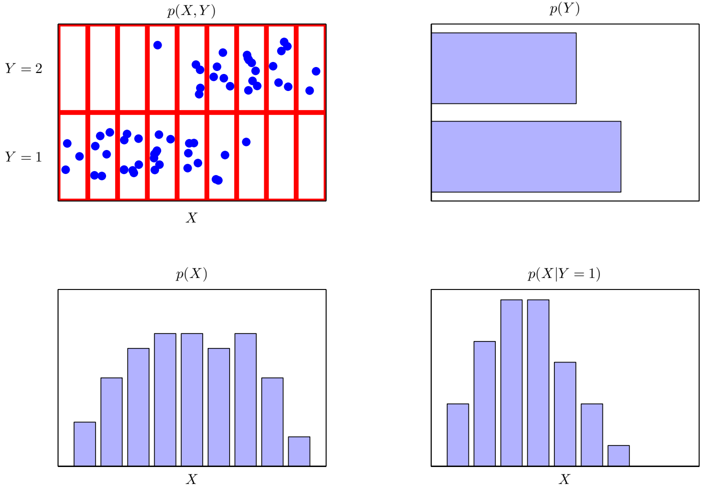
$$\begin{aligned} P(X = x_i, Y = y_j) &= \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \times \frac{c_i}{N} \\ &= P(Y = y_j | X = x_i) P(X = x_i) \end{aligned}$$

Επανάληψη στην θεωρία πιθανοτήτων

Ο κανόνας αθροίσματος $\quad P(X) = \sum_{Y} P(X, Y)$

Κανόνας γινομένου $\quad P(X, Y) = P(Y|X)P(X) = P(X|Y)P(Y)$

Επανάληψη στην θεωρία πιθανοτήτων



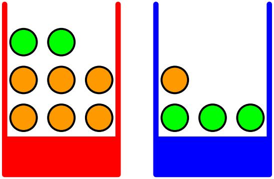
Επανάληψη στην θεωρία πιθανοτήτων

Θεώρημα Bayes (Bayes' Theorem)

$$\begin{aligned} P(X|Y) &= \frac{P(Y|X)P(X)}{P(Y)} \\ \text{Posterior} &= \frac{\text{Likelihood} \times \text{Prior}}{\text{Normalizing Constant}} \end{aligned}$$

- Prior $P(X)$: Εκφράζει την εκ των προτερών πίστη/βεβαιότητα για το ποια είναι η τιμή της X
- Likelihood $P(Y|X)$: Η πιθανότητα κάποιας παρατηρούμενης πληροφορίας (δεδομένα!)
- Posterior $P(X|Y)$: Εκφράζει την εκ των υστέρων πίστη/βεβαιότητα μας (δηλ. μετά την παρατήρηση των δεδομένων) για το ποια είναι η τιμή της X
- Normalizing Constant $P(Y)$: $P(Y) = \sum_X P(Y|X)P(X)$, απλά κανονικοποιεί την posterior ώστε $\sum_X P(X|Y) = 1$

Επανάληψη στην θεωρία πιθανοτήτων



'Αν επιλέξουμε ένα πορτοκάλι, τότε ποιο ήταν το κουτί;'

Τι γνωρίζουμε:

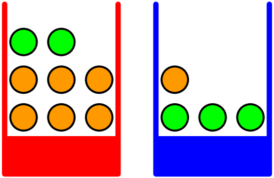
- Επιλογή του κουτιού (τυχαία μεταβλητή X):

$$P(X = blue) = \frac{6}{10}, \quad P(X = red) = \frac{4}{10}$$

- Επιλογή του φρούτου (τ.μ. Y) δοθέντος του κουτιού:

$$\begin{aligned} P(Y = orange|X = blue) &= \frac{1}{4}, \quad P(Y = apple|X = blue) = \frac{3}{4} \\ P(Y = orange|X = red) &= \frac{3}{4}, \quad P(Y = apple|X = red) = \frac{1}{4} \end{aligned}$$

Επανάληψη στην θεωρία πιθανοτήτων



'Αν επιλέξουμε ένα πορτοκάλι, τότε ποιο ήταν το κουτί;'

Τι ψάχνουμε:

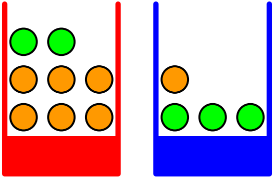
- Τις πιθανότητες

$$P(X = red|Y = orange) \text{ και } P(X = blue|Y = orange)$$

προφανώς αρκεί να βρούμε την $P(X = red|Y = orange)$ αφού ισχύει

$$P(X = blue|Y = orange) + P(X = red|Y = orange) = 1$$

Επανάληψη στην θεωρία πιθανοτήτων



‘Αν επιλέξουμε ένα πορτοκάλι, τότε ποιο ήταν το κουτί;’

Πιθανότητα να επιλεγεί ένα πορτοκάλι

$$P(Y = orange) = P(Y = orange|X = red)P(X = red) + P(Y = orange|X = blue)P(X = blue$$

$$P(Y = orange) = \frac{3}{4}\frac{4}{10} + \frac{1}{4}\frac{6}{10} = \frac{9}{20}$$

Θεώρημα Bayes

$$\begin{aligned} P(X = red|Y = orange) &= \frac{P(Y = orange|X = red)P(X = red)}{P(Y = orange)} \\ &= \frac{1}{4}\frac{6}{10}\frac{20}{9} = \frac{2}{3} \end{aligned}$$

Θα θέλαμε να ορίσουμε πιθανότητες για συνεχείς τυχαίες μεταβλητές.

- Μια συνεχής τυχαία μεταβλητή παίρνει τιμές σε όλο το $\mathbb{R}$ ή σε κάποιο υποσύνολο του
- Υπάρχουν άπειρες και μη αριθμήσιμες τιμές που μπορεί να παίρνει μια συνεχής τυχαία μεταβλητή
  - Για ανάθεση τιμών πιθανότητων δεν μπορούμε να βασιστούμε στον τρόπο που χρησιμοποιήσαμε για διακριτές τυχαίες μεταβλητές
- Αναθέτουμε τιμές βάσει μιας συνάρτησης πυκνότητας πιθανότητας (probability density function)

Επανάληψη στην θεωρία πιθανοτήτων

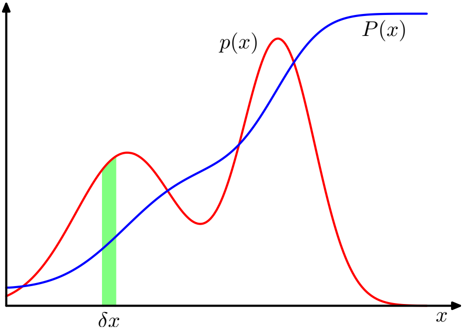
- Έστω συνεχή τυχαία μεταβλητή X. Η συνεχής πιθανοτική κατανομή αναθέτει σε κάθε διάστημα (a, b) του $\mathbb{R}$ την πιθανότητα η τιμή της X να βρίσκεται στο (a, b) βάσει

$$P(x \in (a, b)) = \int_a^b p(x)dx$$

όπου $p(x) = p(X = x) =$ ονομάζεται συνάρτησης πυκνότητας:

$$\int p(x)dx = 1, \quad p(x) \geq 0$$

Επανάληψη στην θεωρία πιθανοτήτων



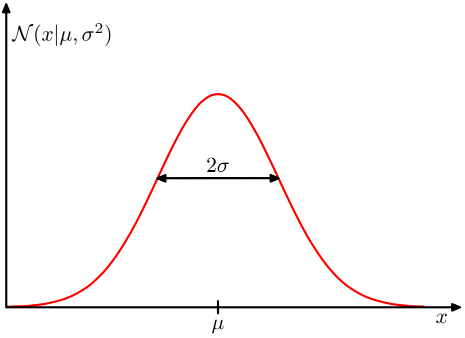
- Η κόκκινη γραμμή δείχνει μια συνάρτηση πυκνότητας πιθανότητας:

$$\int p(x)dx = 1, \quad p(x) \geq 0$$

- Η μπλε δείχνει γραμμή την cumulative distribution function

$$P(z) = P(x \in (-\infty, z)) = \int_{-\infty}^{z} p(x)dx$$

Επανάληψη στην θεωρία πιθανοτήτων



$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

Είναι μακράν η πιο σημαντική κατανομή.

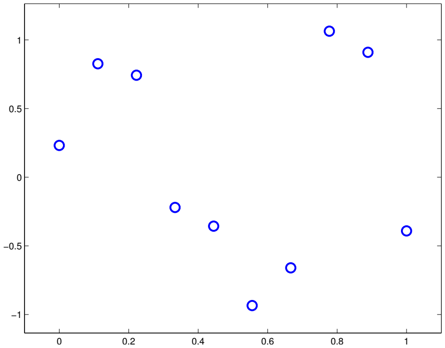
Πιθανοτικό μοντέλο για παλινδρόμηση

- Έστω ότι έχουμε τα ακόλουθα δεδομένα εκπαίδευσης

$$\mathcal{D} = \{\mathbf{x}_n, t_n\}_{n=1}^{N}, \quad t_n \in \mathbb{R}$$

  όπου κάθε $\mathbf{x}_n$ είναι ένα δεδομένο εισόδου και t_n το αντίστοιχο δεδομένο εξόδου
- Πρόβλημα μάθησης: Κατασκευή ενός συστήματος που να μαθαίνει να προβλέπει την έξοδο t_* για κάθε άγνωστο δεδομένο εισόδου $\mathbf{x}_*$

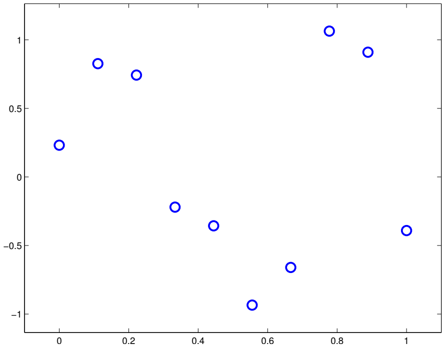
Πιθανοτικό μοντέλο για παλινδρόμηση



Θα θέλαμε να κατασκευάσουμε ένα πιθανοτικό μοντέλο που

- να μαθαίνει μια (ντετερμινιστική) συνάρτηση που περιγράφει την δομή των δεδομένων
- να μοντελοποιεί το θόρυβο που υπάρχει στα δεδομένα

Πιθανοτικό μοντέλο για παλινδρόμηση



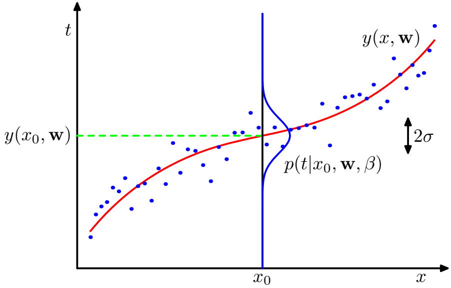
- Δομή: Υποθέτουμε ένα πολυώνυμο

$$y(x, \mathbf{w}) = w_0 + w_1 x + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

- Θόρυβος: Υποθέτουμε ότι ακολουθεί την Gaussian κατανομή

$$t = y(x, \mathbf{w}) + \epsilon, \quad \epsilon \sim \mathcal{N}(\epsilon | 0, \beta^{-1})$$

Πιθανοτικό μοντέλο για παλινδρόμηση



- Οπότε η πιθανοτική κατανομή του δεδομένου εξόδου t_n δοθέντος του δεδομένου εισόδου x_n είναι και αυτή Gaussian

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) = \left(\frac{\beta}{2\pi}\right)^{1/2} \exp\left\{-\frac{\beta}{2}\left(t - y(x, \mathbf{w})\right)^2\right\}$$

Μέθοδος μέγιστης πιθανοφάνειας (maximum likelihood)

Θελούμε να εκτιμήσουμε τις παραμέτρους $(\mathbf{w}, \beta)$ ώστε το μοντέλο να ταιριάξει στα δεδομενα $\Rightarrow$ εκπαίδευση

- Από κοινού κατανομή: Υποθέσουμε ότι κάθε t_n έχει παραχθεί ανεξάρτητα δοθέντος του x_n ώστε

$$\begin{aligned} p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) &= \prod_{n=1}^{N} p(t_n|x_n, \mathbf{w}, \beta) \\ &= p(\mathbf{t}|x, \mathbf{w}, \beta) \end{aligned}$$

όπου $\mathbf{t} = \{t_n\}_{n=1}^{N}$ και $\mathbf{x} = \{x_n\}_{n=1}^{N}$. Η ποσότητα αυτή εξαρτάται (δηλ. η τιμή της μεταβάλλεται!) από τις παραμέτρους $(\mathbf{w}, \beta)$

Μέθοδος μέγιστης πιθανοφάνειας (maximum likelihood)

Θέλουμε να εκτιμήσουμε τις παραμέτρους $(\mathbf{w}, \beta)$

- Μεγιστοποιούμε την από κοινού κατανομή/πιθανότητα των δεδομένων

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n | y(x_n, \mathbf{w}), \beta^{-1})$$

- Λύση

$$\mathbf{w}_{ML} = \arg\max_{\mathbf{w}} \frac{1}{2} \sum_{n=1}^{N} (t_n - y(x_n, \mathbf{w}))^2$$

$$\frac{1}{\beta_{ML}} = \frac{1}{N} \sum_{n=1}^{N} (t_n - y(x_n, \mathbf{w}))^2$$

Επίλογος

- Διάβασμα για το σπίτι: section 1.2 (subsections 1, 2, 4, 5) από το βιβλίο του Bishop
- Επόμενο μάθημα: Γραμμικά μοντέλα παλινδρόμησης και λογιστικής παλινδρόμησης