Στο έγγραφο αυτό συγκεντρώνουμε τις ποιό συχνές παρατηρήσεις που κάναμε στην πρόοδο της μηχανικής μάθησης.

Σημειώστε βέβαια ότι σε όλες τις ερωτήσεις οι μονάδες δεν ήταν "όλες ή μηδέν", αλλά προσμετρήθηκαν κλιμακωτά, ανάλογα την απάντηση.

Γιατί λέμε αφελή τον γκαουσιανό αφελή μπεϋζιανό ταξινομητή;

Τον αποκαλούμε αφελή γιατί κάνει *δύο* απλουστευτικές υποθέσεις: την ανεξαρτησία μεταξύ των χαρακτηριστικών και ότι *όλα τα χαρακτηριστικά ακολουθούν την κανονική (γκαουσιανή) κατανομή.*

Έχετε έναν γκαουσιανό αφελή μπεϋζιανό ταξινομητή και ένα dataset με αριθμητικές τιμές. Στη διαδικασία της εκπαίδευσης, ποια από τις ακόλουθες διαδικασίες πρέπει να ακολουθήσετε:
- Διασταυρούμενη επικύρωση πλέγματος
- Διασταυρούμενη επικύρωση
- Υπερδειγματοληψία
- Καμμία από τις άλλες απαντήσεις

Απαντήθηκε συχνά "Καμμία από τις άλλες απαντήσεις". Ωστόσο, παρότι ο GNB δεν έχει υπερπαραμέτρους, και άρα δεν χρειάζεται διασταυρούμενη επικύρωση πλέγματος, θα πρέπει να υπολογιστούν οι μέσες τιμές και η διακύμανση για όλα τα χαρακτηριστικά, ως προς ένα σύνολο επικύρωσης. Συνεπώς θα εφαρμόσουμε διασταυρούμενη επικύρωση.

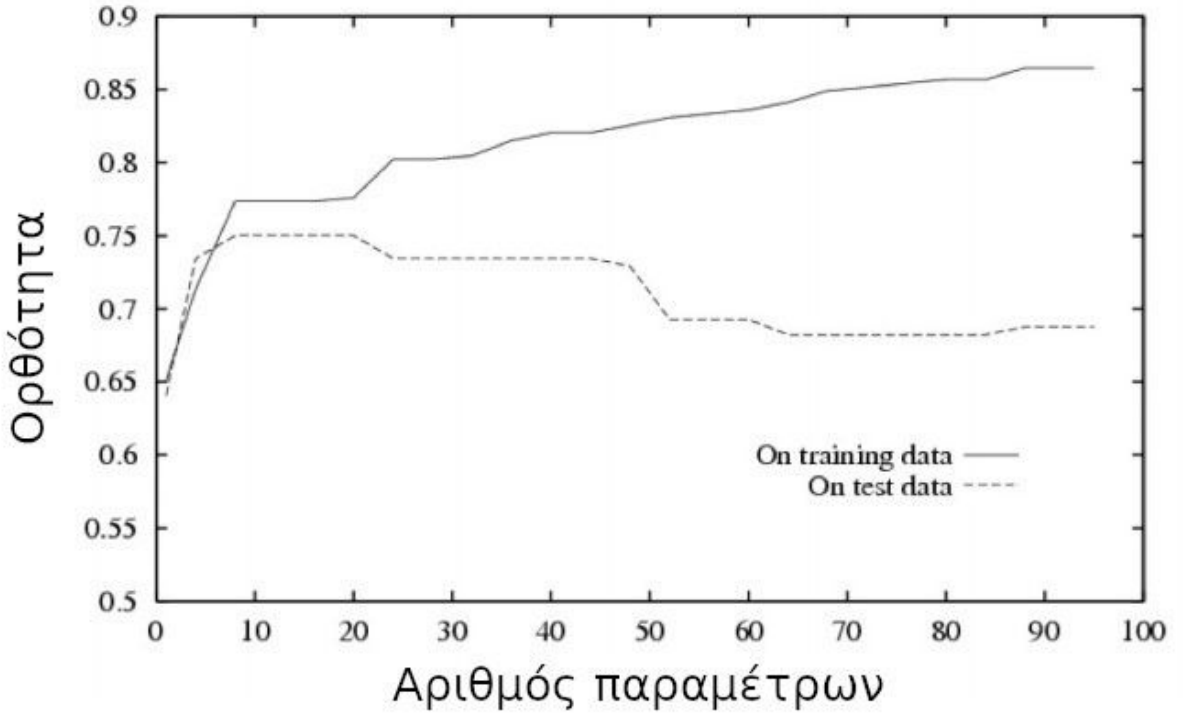
Ένας γιατρός χρησιμοποιεί μηχανική μάθηση (ταξινομητές) για να ανιχνεύσει μια σπάνια ασθένεια σε σύνολα δειγμάτων. Ποια ή ποιες μετρικές αξιολόγησης θα είναι πιο σημαντικό να χρησιμοποιήσει και γιατί;

Η βασική μετρική που τον ενδιαφέρει είναι η ανάκληση (recall) ώστε να μην "χάσει" δείγματα θετικά στην ασθένεια, ακόμα και αν προκύψουν πιθανώς πολλά εσφαλμένα θετικά δείγματα (χειρότερη ακρίβεια - precision). Μπορεί επίσης να χρησιμοποιήσει την F1-macro, ένας αρμονικός μέσος που δεν λαμβάνει υπόψη του το πλήθος των δειγμάτων σε κάθε κλάση.

Εκπαιδεύουμε έναν ταξινομητή αυξάνοντας σταδιακά την πολυπλοκότητά του και φτιάχνουμε τον γράφο της εικόνας που δείχνει την ορθότητα (accuracy) σε σχέση με τον αριθμό των παραμέτρων. Η συνεχής γραμμή είναι για το σύνολο εκπαίδευσης και η διακεκομμένη για το σύνολο ελέγχου. Θεωρούμε ότι ισχύουν οι συνθήκες PAC για τη δειγματοληψία στα σύνολα εκπαίδευσης και ελέγχου. Σημειώστε ότι το σφάλμα ε = 1 - ορθότητα.

Θεωρούμε το αγνωστικό φράγμα PAC το οποίο δίνει τη διαφορά μεταξύ του αληθινού σφάλματος και του σφάλματος εκπαίδευσης για κάθε υπόθεση h από το σύνολο των ταξινομητών που εξετάζουμε.

Με προσεκτική (formal) διατύπωση γράψτε τι εγγυάται το αγνωστικό φράγμα PAC για τις δύο καμπύλες του γράφου.

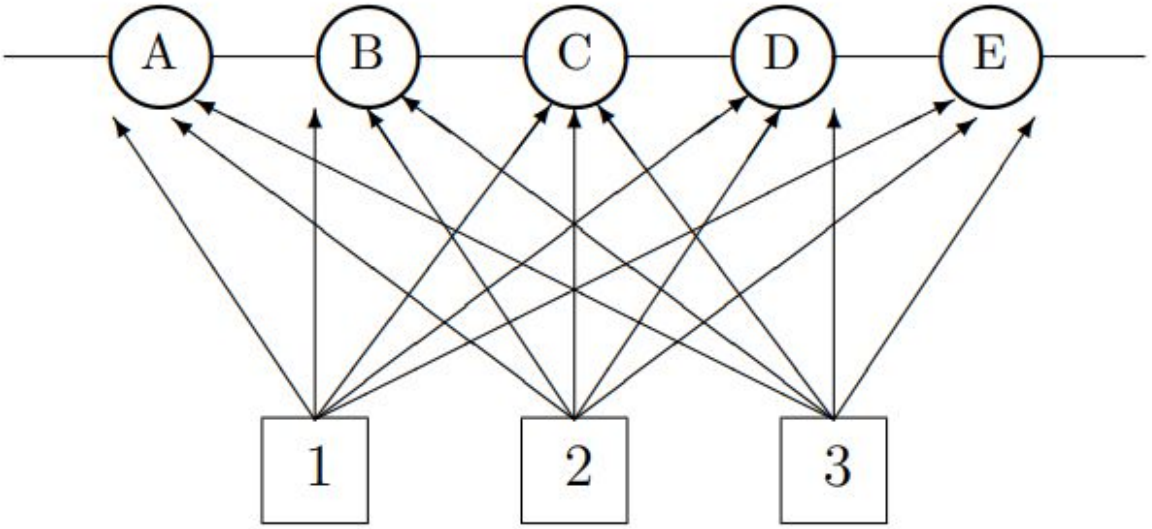


Αν εκπαιδεύσουμε τον ταξινομήτη σε m παραδείγματα, ανεξάρτητα και ομοιόμορφα δειγματοληπτημένα από την κατανομή D(X), τότε, με πιθανότητα (1-δ), η υπερπροσαρμογή (overfitting), δηλαδή η διαφορά μεταξύ της καμπύλης της ορθότητας στο σύνολο εκπαίδευσης και της καμπύλης στο σύνολο ελέγχου, για όλες τις υποθέσεις h του γραφήματος, θα είναι μικρότερη-ίση από ε.
Θεωρούμε ότι η αληθινή ορθότητα είναι η αναμενόμενη τιμή στο σύνολο ελέγχου, για διάφορες τυχαίες δειγματοληψίες συνόλων ελέγχου.

Ένας αλγόριθμος μάθησης βασισμένος στο ξυράφι του Όκαμ αναμένουμε, χωρίς καμία άλλη πληροφορία, να κάνει υποπροσαρμογή ή υπερπροσαρμογή στα δεδομένα; (εξηγείστε εν συντομία)
Εξηγείστε δύο σχέσεις όπου εμφανίζεται το ξυράφι του Όκαμ τις σχέσεις του PAC learning (σε 2-3 προτάσεις).

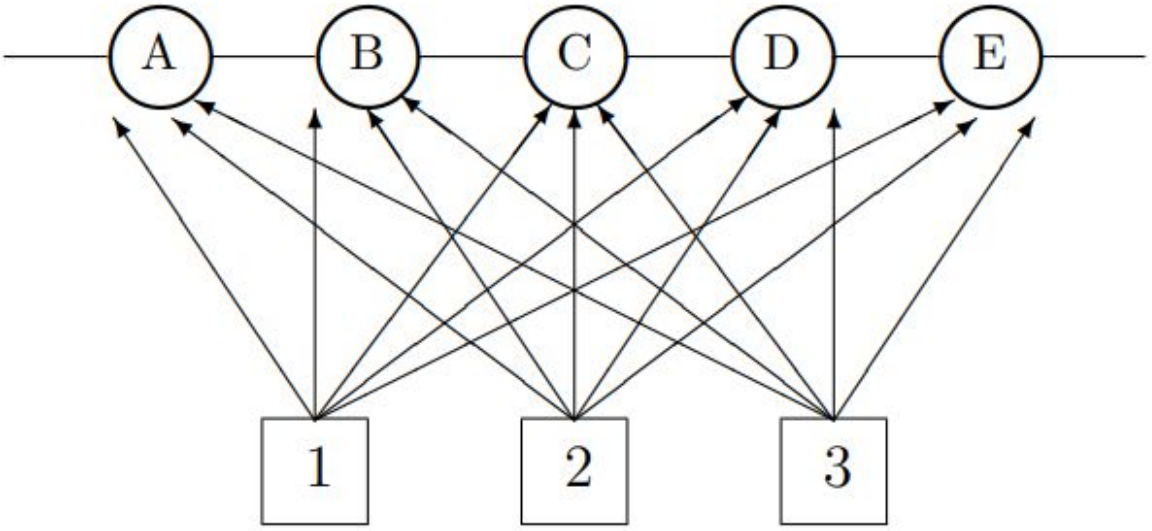
Υποπροσαρμογή: είναι ο απλούστερος της οικογένειάς του, αυτός με την μεγαλύτερη πόλωση.
Το ξυράφι του Όκαμ εμφανίζεται τόσο στη συνεπή όσο και στην ασυνεπή περίπτωση PAC learning, μέσα από το μέγεθος του συνόλου των υποθέσεων |H|.
Γενικά απαντήθηκε το πρώτο σκέλος της ερώτησης, αλλά σχεδόν καθόλου το δεύτερο.

Δίνεται ο αυτο-οργανούμενος χάρτης SOM της εικόνας. Ποιο είναι το πλήθος των προτύπων εισόδου που μπορεί να επεξεργαστεί;



Άπειρο. Η απάντηση 2^3 θα ήταν σωστή αν η ερώτηση ήταν: έστω ότι η είσοδος είναι δυαδική (που δεν είναι απαραίτητο, τα SOM δεν είναι δίκτυα Hopfield). Πόσα *διαφορετικά* πρότυπα εισόδου μπορεί να εμφανιστούν στην είσοδο;

Δίνεται ο αυτο-οργανούμενος χάρτης SOM της εικόνας. Πόσες ομάδες (clusters) μπορεί να ανιχνεύσει;



Το πολύ πέντε. Απαντήθηκε το πέντε, αλλά η ακριβής απάντηση είναι “το πολύ πέντε”, καθώς μπορεί κάποιες συστάδες να συμπίπτουν. Δεν υπήρξε καμία τέτοια απάντηση.

Έστω το σύνολο υποθέσεων H = {h_i(x)=x_i}. Έστω ότι το x αποτελείται από 4 bits, και ότι ε=0.2 και δ=0.05. Πόσα δείγματα χρειαζόμαστε για να μάθουμε αυτό σύνολο υποθέσεων σύμφωνα με το PAC learning (επεξήγηση σε δύο-τρεις προτάσεις);

Οι διαστάσεις του χώρου εισόδου είναι βέβαια 2^4. Ωστόσο, ο οι υποθέσεις μας είναι μόνο 4 το πλήθος: οι x1, x2, x3 και x4, ή αλλιώς οι άξονες των συντεταγμένων του χώρου, ή αλλιώς 4 ταξινομητές που αποφασίζουν με βάση την τιμή ενός και μόνο bit από τα τέσσερα. Συνεπώς m >= 1/0.2 (ln(4) + ln(1/0.05)) = 21,905.

Υπολογίστε, με αιτιολόγηση σε δύο γραμμές, την διάσταση VC για το σύνολο συναρτήσεων της εικόνας, όπου x = [0, 1]

$$F' = \{f : \chi \mapsto \{0,1\},\ f(x) = 1_{t_1 \leq x < t_2} \text{ or } f(x) = 1 - 1_{t_1 \leq x < t_2}, t_1 < t_2 \in [0,1]\}$$

Έστω x1, x2, x3 σε αύξουσα σειρά. Με την F' μπορούμε να αναθέσουμε την ετικέτα "1" στο x1 και στο x3, και την ετικέτα "0" στο x2. Συνεπώς VC(F')>=3.
Αν τώρα λάβουμε τέσσερα σημεία x1, x2, x3, x4 πάντα σε αύξουσα σειρά, δεν μπορούμε, για παράδειγμα, να αναθέσουμε την ετικέτα "1" στο x1 και στο x3 και την "0" στα x2 και x4. Συνεπώς VC(F')=3.