Spam Classification

Παράδειγμα feature vector:

$x(M) =$($\mathbf{1}$(το M έχει το σύμβολο \$), sgn (το M είναι ορθογραφικά σωστό),
links στο M, $\mathbf{1}$(το M έχει τη λέξη "password") , ….)

- Στόχος: Βρές κανόνα $F: \mathbb{R}^n \to \{\pm 1\}$ που αποφασίζει σωστά για **νέα** emails.
- $(x_1, y_1), \dots, (x_m, y_m)$ είναι το **training set**
- Πώς βρίσκεις τον F;

  1) Αποφασίζεις μια οικογένεια κανόνων, από την οποία θα επιλέξεις.

  Εμείς θα μελετήσουμε την: $F_w(x) = \mathrm{sgn}(w^T x) = \mathrm{sgn}\left(\sum_{j=1}^n w_j x_j\right), w \in \mathbb{R}^n$

  Γεωμετρικά: διαχωριστικό υπερεπίπεδο

Spam Classification

2) Πώς βρίσκεις ένα καλό w;

- Ιδανικά: $\forall i = 1, \dots, m, \quad y_i w^T x_i > 0$
- Μπορεί να είναι αδύνατο. Βρές το w που κάνει τα λιγότερα λάθη:

$$min_w \, L(w), \quad L(w) = \frac{1}{m} \sum_{i=1}^{m} \underbrace{\mathbf{1}(y_i w^T x_i \leq 0)}_{f_i(w) = \varphi(y_i w^T x_i),}$$
$$\varphi(z) = \mathbf{1}(z \leq 0)$$

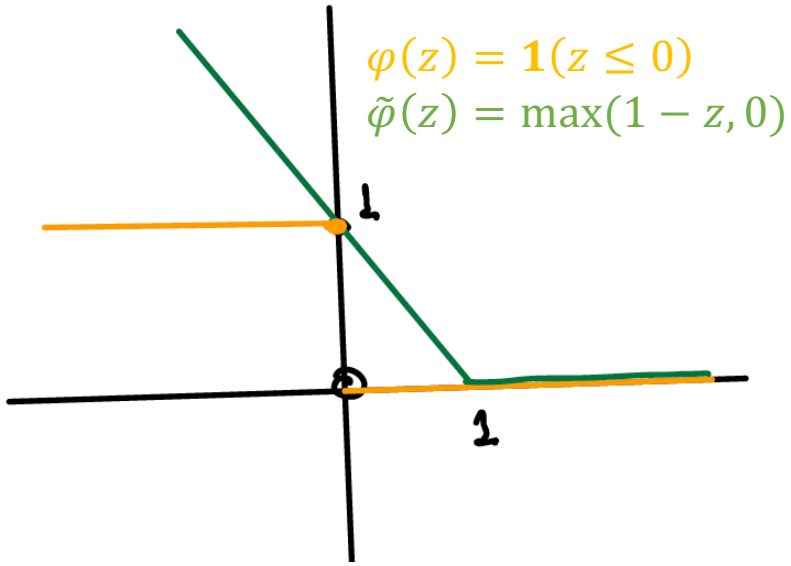
- Αν η φ ήταν κυρτή, τότε η $L(w)$ θα ήταν κυρτή;
- Η φ δεν είναι κυρτή και το πρόβλημα $min_w L(w)$ είναι NP-hard.

Spam Classification

- Λύση: αντικατάστησε τη φ με μία κυρτή που της μοιάζει.

- Νέα, κυρτή συνάρτηση κόστους: $\tilde{\mathrm{L}}(w)$

- $\mathrm{L(w)} \leq \tilde{\mathrm{L}}(w)$



- Για ταχύτερη βελτιστοποίηση: smoothed versions της $\tilde{\varphi}$.