Machine learning lecture slides

COMS 4771 Fall 2020

Classification II: Margins and SVMs

Outline

- Perceptron
- Margins
- Support vector machines
- Soft-margin SVM

Perceptron (1)

- *Perceptron*: a variant of SGD
  - Uses *hinge loss*: $\ell_{\text{hinge}}(s) := \max\{0, 1 - s\}$
  - Uses *conservative updates*: only update when there is classification mistake
  - Step size $\eta = 1$
  - Continues updating until all training examples correctly classified by current linear classifier

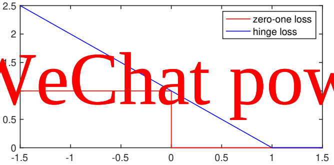


Figure 1: Comparing hinge loss and zero-one loss

Perceptron (2)

- Start with $w^{(0)} = 0$.
- For $t = 1, 2, \ldots$ until all training examples correctly classified by current linear classifier:
  - Pick a training example—call it (x_t, y_t)—misclassified by $w^{(t-1)}$.
  - Update:

$$w^{(t)} := w^{(t-1)} - \nabla \ell_{\text{hinge}}(y_t x_t^{\top} w^{(t-1)}).$$

- Note that whenever $y_t x_t^\top w^{(t-1)} \leq 0$,

$$\nabla \ell_{\text{hinge}}(y_t x_t^\top w^{(t-1)}) = \ell'_{\text{hinge}}(y_t x_t^\top w^{(t-1)}) \cdot y_t x_t = -1 \cdot y_t x_t.$$

  - So update is

$$w^{(t)} := w^{(t-1)} + y_t x_t.$$

- Final solution is of the form

$$\hat{w} = \sum_{i \in S} y_i x_i$$

for some multiset S of $\{1, \ldots, n\}$.

  - Possible to include same example index multiple times in S

Properties of Perceptron

- Suppose $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \{-1, +1\}$ is linearly separable.
- Does Perceptron find a linear separator? (Yes.) How quickly?
- Depends on *margin* achievable on the data set—how much wiggle room there is for linear separators.

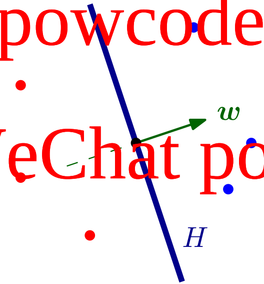


Figure 2: Linearly separable data

Margins (1)

- *Margin* achieved by w on i-th training example is the distance from $y_i x_i$ to decision boundary:

$$\gamma_i(w) := \frac{y_i x_i^\top w}{\|w\|_2}.$$

- *Maximum margin* achievable on all training examples:

$$\gamma_\star := \max_{w \in \mathbb{R}^d} \min_i \gamma_i(w).$$

- **Theorem**: If training data is linearly separable, Perceptron finds a linear separator after making at most $(L/\gamma_\star)^2$ updates, where $L = \max_i \|x_i\|_2$.

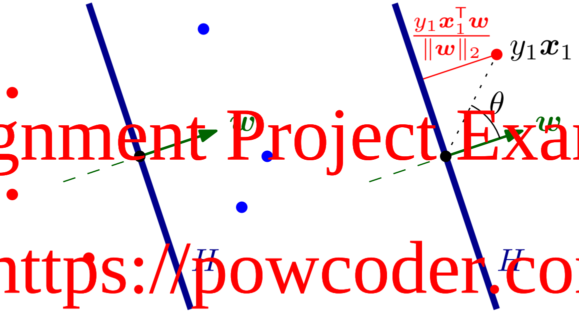


Figure 3: Margins

- Let w be a linear separator:

$$y_i x_i^\mathsf{T} w > 0, \quad i = 1, \ldots, n.$$

- Note: Scaling of w does not change margin achieved on i-th example

$$\gamma_i(w) = \frac{y_i x_i^\mathsf{T} w}{\|w\|_2}.$$

- WLOG assume $y_1 x_1^\mathsf{T} w = \min_i y_i x_i^\mathsf{T} w$.
- So x_1 is closest to decision boundary among all training examples.

- Rescale w so that $y_1 x_1^\mathsf{T} w = 1$.
- Distance from $y_1 x_1$ to decision boundary is $1/\|w\|_2$.
- The shortest w satisfying

$$y_i x_i^\mathsf{T} w \geq 1, \quad i = 1, \ldots, n$$

gives the linear separator with the *maximum margin* on all training examples.

- Weight vector of maximum margin linear separator: defined as solution to optimization problem

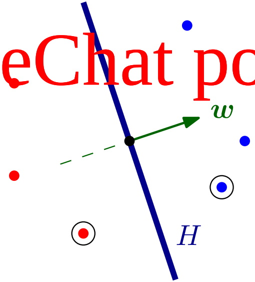
$$\min_{w \in \mathbb{R}^d} \quad \frac{1}{2}\|w\|_2^2$$

$$\text{subject to} \quad y_i x_i^{\mathsf{T}} w \geq 1, \quad i = 1, \ldots, n.$$

(The 1/2 prefactor is customary but inconsequential.)

- This is the *support vector machine (SVM)* optimization problem.
- Feasible when data are linearly separable.
- Note: Preference for the weight vector achieving the maximum margin is another example of inductive bias.

- Just like least norm solution to normal equations (and ridge regression), solution w to SVM problem can be written as $\sum_{i=1}^{n} \alpha_i y_i x_i$ for some $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ (in fact, $\alpha_i \geq 0$)
  - (Adding $r \in \mathbb{R}^d$ orthogonal to span of x_i's to weight vector can only increase the length without changing the constraint values.)
- The examples (x_i, y_i) for which $\alpha_i \neq 0$ are called *support vector examples*: they have $y_i x_i^{\top} w = 1$ and are closest to decision boundary.

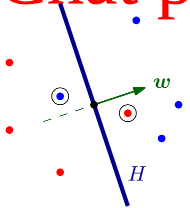


- What if not linearly separable? SVM problem has no solution.
- Introduce *slack variables* for constraints, and $C \geq 0$:

$$\min_{w \in \mathbb{R}^d, \xi_1, \ldots, \xi_n \geq 0} \quad \frac{1}{2}\|w\|_2^2 + C \sum_{i=1}^{n} \xi_i$$

$$\text{subject to} \quad y_i x_i^{\mathsf{T}} w \geq 1 - \xi_i, \quad i = 1, \ldots, n.$$

- This is the *soft-margin SVM* optimization problem.
  - A *constrained* convex optimization problem
- For given w, $\xi_i / \|w\|_2$ is distance that x_i has to move to satisfy $y_i x_i^{\mathsf{T}} w \geq 1$.



- Equivalent unconstrained form:

$$\min_{w \in \mathbb{R}^d} \quad \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \max\{0, 1 - y_i x_i^{\mathsf{T}} w\}$$

- Rewriting using $\lambda = 1/(nC)$ and ℓ_{hinge}:

$$\min_{w \in \mathbb{R}^d} \quad \frac{1}{n}\sum_{i=1}^{n} \ell_{\text{hinge}}(y_i x_i^{\mathsf{T}} w) + \frac{\lambda}{2}\|w\|_2^2.$$

- Same template as ridge regression, Lasso, . . . !
  - Data fitting term (using a surrogate loss function)
  - Regularizer that promotes inductive bias
  - λ controls trade-off of concerns
- Both SVM and soft-margin SVM can be kernelized