NONLINEAR METHODS AND KERNELS

Matthieu R Bloch

January 31, 2019

LOGISTICS

Lecture slides and notes

- Lecture 6 and Lecture 4 notes updated
- Report typos and errors on Piazza (thank you!)

Self assignment graded

- Grades released later today
- Check the solutions and try to understand where you made mistaktes (if any)
- You *need* to master conditional probabilities and expectations

Problem set #1 assigned

- Due Friday *Feb 8, 2019 11:59pm EST* for on-site students
- Due Friday *Feb 15, 2019 11:59pm EST* for DL students
- Hard deadlines two days after your deadline
- Bonuses for -ing and on-time submission

RECAP: PLA

Consider $\mathbf{x} = [1\, x_1, \cdots, x_d]^\intercal \in \mathbb{R}^{d+1}$ and hyperplane of the form $\theta^\intercal \mathbf{x} = 0$.

Data $\mathcal{D} \triangleq \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ such that $y_i \in \{\pm 1\}$

Perceptron Learning Algorithm (PLA) invented by Rosenblatt in 1958 to find separating hyperplanes

- Start from guess for $\theta^{(0)}$ and go over data points in sequence to update

$$\theta^{(j+1)} = \begin{cases} \theta^{(j)} + y_i \mathbf{x}_i \text{ if } y_i \neq \operatorname{sgn}\left({\theta^{(j)}}^\intercal \mathbf{x}_i\right) \\ \theta^{(j)} \text{ else} \end{cases}$$

- Geometric intuition behind operation
- Stochastic gradient descent view

PLA finds a *non parametric* linear classifier

- Can be viewed as single layer NN

Theorem.

PLA finds a separating hyperplane if the data is linearly separable

MAXIMUM MARGIN HYPERPLANE

“All separating hyperplanes are equal but some are more equal than others”

Margin $\rho(\mathbf{w}, b) \triangleq \min_i \frac{|\mathbf{w}^\intercal \mathbf{x}_i + b|}{\|\mathbf{w}\|_2}$

The *maximum margin hyperplane* is the solution of

$$(\mathbf{w}^*, b^*) = \operatorname*{argmax}_{\mathbf{w},b} \rho(\mathbf{w}, b)$$

- Larger margin leads to better generalization

Definition.

The canonical form $(\mathbf{w}, b)$ of a separating plane is such that

$$\forall i\ y_i(\mathbf{w}^\intercal \mathbf{x}_i + b) \geq 1 \text{ and } \exists i^* \text{ s.t. } y_{i^*}(\mathbf{w}^\intercal \mathbf{x}_{i^*} + b) = 1$$

For canonical hyperplanes, the optimization problem is

$$\operatorname*{argmin}_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2 \text{ s.t. } \forall i \quad y_i(\mathbf{w}^\intercal \mathbf{x}_i + b) \geq 1$$

- this is a constrained quadratic program
- we know how to solve this really well
- Will come back when we talk about *support vector machines*

OPTIMAL SOFT-MARGIN HYPERPLANE

What if our data is not linearly separable?

- The constraint $\forall i \quad y_i(\mathbf{w}^\intercal \mathbf{x}_i + b) \geq 1$ cannot be satisfied
- Introduce slack variables $\xi_i > 0$ such that $\forall i \quad y_i(\mathbf{w}^\intercal \mathbf{x}_i + b) \geq 1 - \xi_i$

The optimal soft-margin hyperplane is the solution of the following

$$\underset{\mathbf{w},b,\boldsymbol{\xi}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{w}\|_2^2 + \frac{C}{N}\sum_{i=1}^{N}\xi_i \text{ s.t. } \forall i \quad y_i(\mathbf{w}^\intercal \mathbf{x}_i + b) \geq 1 - \xi_i \text{ and } \xi_i \geq 0$$

$C > 0$ is a cost set by the user, which controls the influence of outliers

NON LINEAR FEATURES

LDA, logistic, PLA, are all *linear classifiers*: classification region boundaries are *hyperplanes*

- Some datasets are not linearly separable!

We can create *nonlinear* classifiers by *transforming* the data through a non linear map $\Phi : \mathbb{R}^d \to \mathbb{R}^p$

$$\Phi : \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} \to \begin{bmatrix} \phi_1(\mathbf{x}) \\ \vdots \\ \vdots \\ \phi_p(\mathbf{x}) \end{bmatrix}$$

One can then apply linear methods on the transformed feature vector $\Phi(\mathbf{x})$

Example.

Ring data

Challenges: if $p \gg n$ this gets computationally challenging and there is a risk of overfitting!