Lec 14

Thursday, October 31, 2019 11:09

Recap:

Supervised learning using ensembles

General idea: Combine many bad models (regression or classifiers) to make one good one.

Last time: Bagging

Combine many low-bias hi-variance models to make a low-bias-low-var model

Today: Boosting

Idea: Combine many weak learners (high bias) to make a strong learner

A weak learner a supervised learning algo that does slightly than guessing.

Suppose we have a classification model f

Want to train a new model

$$f + \beta G$$

Where G is some weak learner

β is some weight

$$\underset{\beta, G}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \text{loss}\left(y_i,\ f(x_i) + \beta G(x_i)\right) \right.$$

$$= \sum_{i=1}^{n} \exp\left\{ -y_i \left(f(x_i) + \beta G(x_i) \right) \right\}$$

$$= \sum_{i=1}^{n} \exp\left(-y_i f(x_i)\right) \exp\left(-\beta y_i G(x_i)\right)$$

$$= \sum_{i=1}^{n} W_i \exp\left(-\beta y_i G(x_i)\right)$$

$W_i = e^{-y_i f(x_i)}$

for any β, $\rightarrow$ is a sort of weighted loss on G

Let G be some learner on the weighted dataset

$$\left\{ W_i \rightarrow (x_i, y_i) \right\}_{i=1}^{n}$$

(doesn't need to use exp loss; can be CART stump, logistic reg'n)

With this G fixed, consider the optimal β:

$\text{sign}\, G(x_i) \in \{\pm 1\}$

$$\min_{\beta} \left\{ \sum_{i=1}^{n} W_i \exp\left(-\beta y_i \cancel{G(x_i)}\right) \right.$$

$$= e^{-\beta} \left(\sum_{i: y_i = G(x_i)} W_i \right) + e^{\beta} \left(\sum_{i: y_i \neq G(x_i)} W_i \right)$$

Solve for β:

$$\beta = \frac{1}{2} \log\left(\frac{1 - \text{err}}{\text{err}} \right)$$

$$err = \frac{\sum_i w_i \, \mathbb{I}[y_i \neq \text{sign } G(x_i)]}{\sum_i w_i}$$

↑ weighted avg misclassification loss of G

New model: $f(x) + \beta G(x)$

Ada Boost Algo:

Let $f_1 =$

iterate.

(More explicitly: see slides)

SVM (Support Vector Machine)

Motivation: Which linear seperator is the best?

Idea: Want max margin

Consider the hyperplane $\beta^T x + \beta_0$

Classify as + when $\beta^T x + \beta_0 > 0$

as − when $\beta^T x + \beta_0 < 0$

Our classes are $y_i \in \{+1, -1\}$

$$y_i \cdot (\beta^T x_i + \beta_0) = \begin{cases} > 0 & \text{When classification is correct} \\ < 0 & \text{--//-- is incorrect} \end{cases}$$

That the hyperplane properly separates the data is

$$y_i (\beta^T x_i + \beta_0) > 0 \quad \forall i$$

How big is the margin?

$y_i (\beta^T x_i + \beta_0)$ = How many steps of length $\|\beta\|_2$ in direction $+\beta$ or $-\beta$ we need to take to go from x to hyperplane

$$\Rightarrow \text{distance to the hyperplane} = \frac{y_i (\beta^T x_i + \beta_0)}{\|\beta\|_2}$$

$$\text{Margin} = \min_{i=1,\cdots,n} \frac{y_i (\beta^T x_i + \beta_0)}{\|\beta\|_2}$$

= distance btw hyperplane & nearest pt

Max margin hyperplane:

$$\max_{\beta, \beta_0} \min_{i=1,\cdots,n} \frac{y_i (\beta^T x_i + \beta_0)}{\|\beta\|_2}$$

$$= \max_{\beta, \beta_0, M} M$$

$$\|\beta\|_2 M \le y_i(\beta^T x_i + \beta_0) \quad \forall i$$

$$= \max_{\beta, \beta_0} 1/\|\beta\|_2$$

$$\text{s.t.} \quad y_i(\beta^T x_i + \beta_0) \ge 1 \quad i = 1, \dots, n$$

$$= \min_{\beta, \beta_0} \|\beta\|_2^2$$

$$\text{s.t.} \quad y_i(\beta^T x_i + \beta_0) \ge 1 \quad \forall i$$