CS281B/Stat241B (Spring 2006) Statistical Learning Theory **Lecture: 11**

Ensemble Methods

Lecturer: Peter Bartlett *Scribe: Junming Yin*

1 Introduction

The idea of ensemble methods is to learning a committee of of base classifiers and then predict new data point by combing the prediction of individual classifier. Boosting (especially AdaBoost) and bagging are two typical examples of ensemble methods. Figure 1 sketches the basic idea of ensemble methods.

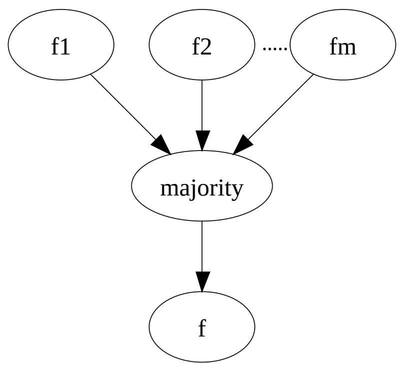

Figure 1: Sketch of ensemble methods

2 Examples of base classifiers

- (i) **Decision trees**: decision trees classify data by sorting them down the tree from the root to leaf node. Each node in the tree is labeled with a decision rule (such as $x_i \geq 0$) and each branch descending from that node corresponds to one possible value for this decision (such as true or false). The data point is finally classified as the label in leaf node with $y = \{+1, -1\}$. Figure 2 shows an example of decision tree with depth 3.
- (ii) **Decision bumps**: decision bumps are simply trees with 1 root node. For example, we classify the data point by looking at whether $(x_i < \theta)$ is true or not at the root node. For this class of base classifiers, we can minimize the empirical risk $\hat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} l(f(x_i,), y_i)$ by enumerating each dimension of the data, which takes time $O(dn \log n)$ where d is the dimension of the data.
- (iii) **Linear threshold functions**: in this case, the decision boundary is a hyperplane. We can also combine them in a weighted way to get a form of neural network:

$$f(x) = sign(\sum_{i=1}^{m} \alpha_i f_i(x)) = sign(\sum_{i=1}^{m} \alpha_i sign(\theta_i^T x - \theta_{i0}))$$

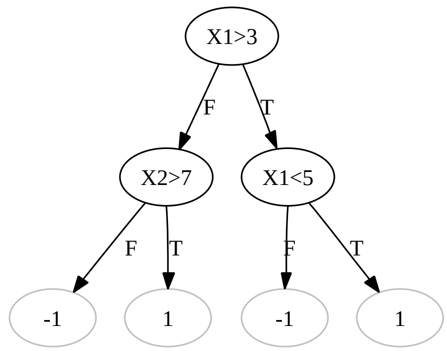

Figure 2: Decision tree

3 AdaBoost

The procedure of AdaBoost is summarized as follows and the details of the algorithm is shown in Figure 3.

- Maintain weighting D_t over the training data $\{1, \ldots, n\}$.
- Chooses f_t sequentially to minimize the D_t-weighted empirical risk.
- Updates D_t to D_{t+1} to emphasize the mistakes.

INITIALIZATION: $D_1(i) = \frac{1}{n}$ for all $1, \ldots, n$ and $F_0(x) = 0$ for all x.
UPDATE: for $t = 1, \ldots, T$

- choose f_t from the family of base classifiers $\mathcal{G}$ to minimize the weighted error $\epsilon_t = \sum_{i=1}^n D_t(i) 1[f_t(x_i) \neq y_i]$
- update $F_t = F_{t-1} + \alpha_t f_t$ where $\alpha_t = \frac{1}{2}\log(\frac{1-\epsilon_t}{\epsilon_t})$.
- update the weights: $D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times e^{-\alpha_t f_t(x_i) y_i}$ where Z_t is a normalization term to ensure D_{t+1} is a distribution over the training data.

Figure 3: AdaBoost algorithm.

Theorem 3.1. We have:

$$\hat{L}(F_T) = \frac{1}{n}\sum_{i=1}^n 1[y_i F_T(x_i) \leq 0] \leq \prod_{i=1}^T 2\sqrt{\epsilon_t(1-\epsilon_t)}$$

Furthermore,

$$\hat{L}(F_T) \leq (1 - 4\gamma^2)^{T/2}$$

if there exists a positive γ such that $\epsilon_t \leq \frac{1}{2} - \gamma$ for all t.
PROOF. For brevity, we suppress the index T of function F_T.

$$y_i F(X_i) \leq 0 \Longleftrightarrow \exp(-y_i F(X_i)) \geq 1 \Longleftrightarrow \exp(-y_i \sum_{t=1}^T \alpha_t f_t(x_i)) \geq 1$$

Hence,

$$\begin{aligned}
\hat{L}(F) &\leq \frac{1}{n}\sum_{i=1}^{n}\exp\left(-y_i\sum_{t=1}^{T}\alpha_t f_t(x_i)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\prod_{t=1}^{T}\exp\left(-y_i\alpha_t f_t(x_i)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\prod_{t=1}^{T} Z_t\frac{D_{t+1}(i)}{D_t(i)}\left(\text{since}\,\frac{D_{t+1}(i)}{D_t(i)} = \frac{\exp(-y_i\alpha_t f_t(x_i))}{Z_t}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\prod_{t=1}^{T} Z_t\right)\frac{D_{T+1}(i)}{D_1(i)} \\
&= \left(\prod_{t=1}^{T} Z_t\right)\left(\sum_{i=1}^{n} D_{T+1}(i)\right)\left(\text{since}\, D_1(i) = \frac{1}{n}\right)
\end{aligned}$$

If we choose Z_t such that $\sum_{i=1}^{n} D_{T+1}(i) = 1$. Then

$$\hat{L}(F) \leq \prod_{t=1}^{T} Z_t$$

We'd like to choose α_t to minimize the normalization factor:

$$\begin{aligned}
Z_t &= \sum_{i:y_i=f_t(i)} D_t(i)e^{-\alpha_t} + \sum_{i:y_i\neq f_t(i)} D_t(i)e^{\alpha_t} \\
&= (1-\epsilon_t)e^{-\alpha_t} + \epsilon_t e^{\alpha_t}
\end{aligned}$$

By differentiating with respect to α_t, we have:

$$-e^{-\alpha_t}(1-\epsilon_t) + e^{\alpha_t}\epsilon_t = 0 \Longrightarrow e^{2\alpha_t} = \frac{1-\epsilon_t}{\epsilon_t} \Longrightarrow \alpha_t = \frac{1}{2}\log(\frac{1-\epsilon_t}{\epsilon_t})$$

With this particular choice,

$$Z_t = (1-\epsilon_t)\sqrt{\frac{\epsilon_t}{1-\epsilon_t}} + \epsilon_t\sqrt{\frac{1-\epsilon_t}{\epsilon_t}} = 2\sqrt{\epsilon_t(1-\epsilon_t)}$$

Therefore,

$$\hat{L}(F) \leq \prod_{t=1}^{T} 2\sqrt{\epsilon_t(1-\epsilon_t)}$$

Moreover, if there exists a positive γ satisfying $\epsilon_t \leq \frac{1}{2} - \gamma$, then

$$Z_t \leq 2\sqrt{(\frac{1}{2}-\gamma)(\frac{1}{2}+\gamma)} = 2\sqrt{\frac{1}{4}-\gamma^2} = \sqrt{1-4\gamma^2}$$

which implies $\hat{L}(F) \leq (1-4\gamma^2)^{T/2}$. The first inequality above stems from the fact that the function $f(x) = \sqrt{x(1-x)}$ is concave and uniquely maximized at the point $x = \frac{1}{2}$.

We can extend this result to show large margins on the training data. Define $\bar{F} = \frac{F_T}{\sum_{t=1}^{T}\alpha_t} = \frac{\sum_{t=1}^{T}\alpha_t f_t}{\sum_{t=1}^{T}\alpha_t}$. Notice that $\epsilon_t \leq \frac{1}{2}$ implies $\alpha_t \geq 0$ hence $\bar{F} \in co(\mathcal{G})$.

Theorem 3.2.

$$\frac{1}{n}\sum_{i=1}^{n} 1[y_i\bar{F}(x_i) \leq \gamma] \leq \prod_{i=1}^{T} 2\sqrt{\epsilon_t^{1-\gamma}(1-\epsilon_t)^{1+\gamma}}$$

IF $\epsilon_t \leq \frac{1}{2} - \gamma$ for all t, then RHS decreases to zero exponentially quickly in T.