We have examined the bias variance decomposition of error. The error under consideration was of a set of models on a single test example z_i.

$S' \sim S$

$m' < m$

We learned variants of a model by training it on S' different versions (randomly subsampled) of the training set. Complex models have low bias but are sensitive to S' the particular composition of the training set. The model variants of the complex models differ wildly (overfitting S') and therefore depict high variance in their predictions on a single test example z_i.

Now we shall examine how the averaged predictions of the different versions of the model would compare with the predictions of a single model. In other words, how the error incurred by the averaged prediction of multiple models would compare with the error by a single prediction.

over a distribution

$\mathbb{E}_{S'}$

In our earlier analysis we were taking expectation with respect to the models (different versions) and we were taking average of the error over the limited count of the test examples. $z_1\, z_2 \ldots z_T$ (fixed)

In our further analysis we assume that the test examples are in abundance and the models (different versions) are limited in count.

Therefore we take $\underline{\text{expectation}}$ over the test examples $\mathbb{E}_Z$ Z and we take $\underline{\text{average over}}$ the countable models. $\frac{1}{M}\sum_{i=1}^{M}$

Assume there are M different versions of the model, and their predictions on a given test example Z are given by $\hat{y}_k(Z)$.

The averaged prediction of a committee of such M models on a test example Z is called as the "Committee" prediction

$$y_{com} \qquad \hat{y}_{com}(Z) = \frac{1}{M}\sum_{k=1}^{M} \hat{y}_k(Z) \quad \leftarrow \text{averaged prediction}$$

single test example.

Prediction by an individual k^{th} model is

$$\text{single model} \quad \hat{y}_k(Z) = \underbrace{f(Z)}_{\text{true}} + \underbrace{\epsilon_k(Z)}_{\text{noise}}$$

Expected sum of squares error, by the k^{th} model:

$$\text{single model} \quad \text{Error}_{\text{Individual}} = \mathbb{E}_Z\left[\{\underbrace{\hat{y}_k(Z) - f(Z)}_{\text{error}}\}^2\right] = \mathbb{E}_Z\left[\underbrace{\epsilon_k(Z)^2}_{\text{noise}}\right]$$

Averaged error of the individual models

$$\text{Error}_{\text{AvgInd.}} = \frac{1}{M}\sum_{k=1}^{M}\mathbb{E}_Z\left[\epsilon_k(Z)^2\right]$$

Average of the expected error of individual models.

Expect sum of squares error for committee predictions,

$$\text{Error}_{\text{Committee}} = \mathbb{E}_Z\left[\{\underbrace{\frac{1}{M}\sum_{k=1}^{M}\hat{y}_k(Z)}_{\text{committee's prediction}} - \overbrace{f(Z)}^{\text{true}}\}^2\right]$$

$$\frac{1}{\cancel{M}}\cancel{M} f(\underline{z})$$

$$= \mathbb{E}_{\underline{z}}\left[\left\{\frac{1}{M}\sum_{k=1}^{M}\hat{y}_k(\underline{z}) - \frac{1}{M}\sum_{k=1}^{M} f(\underline{z})\right\}^2\right]$$

$$= \mathbb{E}_{\underline{z}}\left[\left\{\frac{1}{M}\sum_{k=1}^{M}\left(\hat{y}_k(\underline{z}) - f(\underline{z})\right)\right\}^2\right]$$

$$= \mathbb{E}_{\underline{z}}\left[\left\{\left(\frac{1}{M}\right)\sum_{k=1}^{M}\underbrace{\epsilon_k(\underline{z})}_{\text{noise.}}\right\}^2\right]$$

$$= \frac{1}{M^2}\mathbb{E}_{\underline{z}}\left[\left(\epsilon_1(\underline{z}) + \epsilon_2(\underline{z}) + \cdots + \epsilon_M(\underline{z})\right)^2\right]$$

$(a+b+c+d+e)^2 = a^2+b^2+\ldots+ab+bc+ac+\ldots$

$$= \frac{1}{M^2}\mathbb{E}_{\underline{z}}\left[\sum_k \epsilon_k(\underline{z})^2 + \sum_{\substack{k=1 \\ j=1 \\ i\neq j}}^{\substack{i=M \\ j=M}} \epsilon_i(\underline{z})\,\epsilon_j(\underline{z})\right]$$

The expectation operator being linear, can be applied to the individual operands of the summation.

uncorrelated noise

$$= \frac{1}{M^2}\left[\sum_k \underline{\mathbb{E}_{\underline{z}}\left(\epsilon_k(\underline{z})^2\right)} + \sum_{\substack{k \\ j}} \underbrace{\mathbb{E}\left[\epsilon_i(\underline{z})\,\epsilon_j(\underline{z})\right]}_{=0}\right].$$

We assume the noise to be zero mean and uncorrelated

$$\mathbb{E}_{\underline{x}}\left[\epsilon_k(\underline{x})\right] = 0 \qquad \mathbb{E}_{\underline{z}}\left[\epsilon_k(\underline{z})\,\epsilon_l(\underline{z})\right] = 0 \quad k\neq l$$

$$\therefore E_{\text{error}_{\text{Committee}}} = \frac{1}{M^2}\left(\sum_k \underline{\mathbb{E}_{\underline{z}}\left(\epsilon_k(\underline{z})^2\right)}\right).$$

$$\text{Error}_{\text{Committee}} = \frac{1}{M^2} \cdot M \cdot \text{Error}_{\text{AvgInd}}.$$

Test error

$$\boxed{\text{Error}_{\text{com}} = \frac{1}{M} \text{Error}_{\text{AvgInd}}.}$$

$\frac{1}{M}$ (Average of the expected errors of the individual models)

Average error of an individual model can be reduced by a factor of M, simply by averaging M versions of a model forming a committee.

However, this reduction in error comes in if the errors due to the individual models are uncorrelated.

For highly correlated models, the reduction in error is small. But still, $\boxed{\text{Error}_{\text{Committee}} \leq \text{Error}_{\text{AvgInd}}.}$

$S' \subset S$
$m' \leq m$

To bring a significant reduction in error, we must ensure that the individual models have complementary capabilities.

error rate $< \left(\frac{1}{2} - \gamma\right)$ small

Base classfiers

Ada Boost (Adaptive Boosting)

AdaBoost is an algorithm that forms a committee of weak classifiers. It learns a hypothesis (a committee) with a low Training set empirical risk.

Given input is a training set S of m examples

$$S = \{(x_1, y_1)\ (x_2, y_2) \cdots (x_m, y_m)\}$$

For each example, the label $y_i = f(x_i)$ where f is a true labelling function.

Training set S.

$S \rightarrow \boxed{\text{Learner}} \rightarrow$ Model.

$x \rightarrow \boxed{\text{Model}} \rightarrow y$

Training set S.

$S' \sim S$.

$S'_{1,2,3,4 \dots M}$ $\rightarrow \boxed{\text{Learner}} \rightarrow$

$M_1 \rightarrow \hat{y}_1$

$M_2 \rightarrow \hat{y}_2$

M_3

$\vdots$

$M_M \rightarrow \hat{y}_M$

$\frac{1}{M}\sum \hat{y}$

Bagging

Bootstrap Aggregation

$$\hat{y}_{com} = \sum_{k=1}^{M} \hat{y}_k(z)$$

$Error_{com}$

K-fold.

Cross validation

used to adjust $\underbrace{\text{hyper parameters}}$

estimate the generalization performance.

T:

SVMs : C/λ

The Adaboost learner is called as a booster.

The booster works in a sequence of rounds.

At round t, the booster defines a distribution of weights over the training examples.

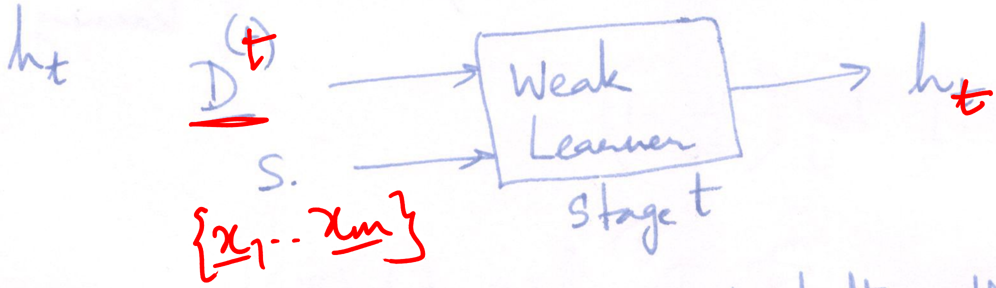
$x_i \to D_i^{(t)}$

Weight on an example x_i is denoted as $D_i^{(t)}$ and with the distribution in stage t is denoted as $D_i^{(t)}$

$$\sum_{i=1}^{m} D_i^{(t)}(t) = 1.$$

i.e. the sum of the probability mass over all the examples is 1.

The weak learner at stage t produces a weak hypothesis h_t



(weak hypothesis) performance just better than random

A weak learner performs just better than random

It has an error rate that is at most $\frac{1}{2} - \gamma$ where γ is a small number.

training set $\boxed{\epsilon_t \leq \frac{1}{2} - \gamma}$

If the hypothesis at stage t, i.e. h_t has an error ϵ_t on the training set, then Adaboost assigns a weight for h_t as $w_t = \frac{1}{2} \log\left(\frac{1}{\epsilon_t} - 1\right)$

A hypothesis h_t with a large error would get a small weight.

The booster also updates the probability mass distribution over the examples after every stage.

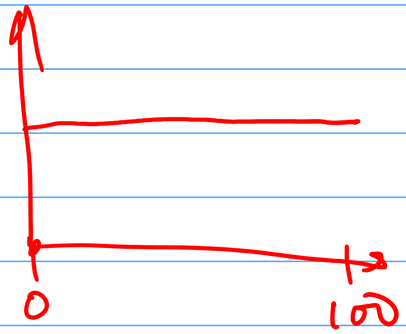
$S = \{ x_1 \quad x_2 \quad x_3 \quad \ldots \}$

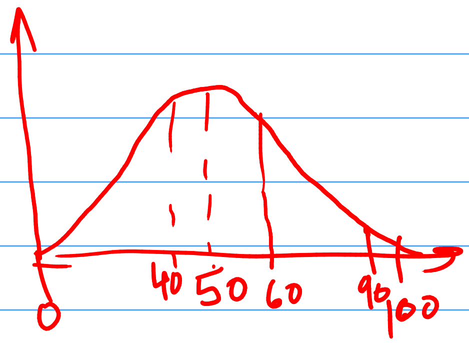
$S' \sim S$

$x_1 \quad x_2$

$x_3 \quad x_4$

How to sample from a distribution?





It $\underline{\text{decreases}}$ the weights of the $\underline{\text{correctly classified}}$ examples and $\underline{\text{increases}}$ the weights of the $\underline{\text{wrongly classified}}$ examples.

$$D_i^{(t+1)} = \frac{D_i^{(t)} \underline{\exp}\left(-w_t y_i h_t(x_i)\right)}{\sum_{j=1}^{m} D_j^{(t)} \exp\left(-w_t y_j h_t(x_j)\right)} \quad \forall i$$

?

wrong: $\exp(+ve) \uparrow$

correct: $\exp(-ve) \downarrow$

$t = 1, 2 \ldots T$

$h_1 \; h_2 \qquad h_T$

The output hypothesis

weighted Committee. prediction

$$\underbrace{h_s(x)}= \operatorname{sign}\left(\sum_{t=1}^{T} w_t h_t(x)\right).$$

where there are T stages of learning.

? where $w_t = \frac{1}{2}\log\left(\frac{1}{\epsilon_t} - 1\right)$ ✓.

indicator function

$= 1$ if cond is satisfied.

and $\epsilon_t = \sum_{i=1}^{m} \underbrace{D_i^{(t)}} \mathbb{1}_{\underbrace{[y_i \neq h_t(x_i)]}_{\text{condition}}}$

It can be showed that $\underbrace{L_S(h_s)}_{\text{empirical risk}} \leq \exp(-2\gamma^2 T)$.

As $T \to \infty$ $L_S(h_s) \to 0$

$h_s(x)$ is a $\underline{\text{halfspace}}$ over the predictions of the weak learners.