Machine learning lecture slides

Assignment Project Exam Help

https://eduassistpro.github.i

Add WeChat edu_assist_pr

Classification IV: Ensemble methods

Assignment Project Exam Help

https://eduassistpro.github.i

Add WeChat edu_assist_pr

Overview

- Bagging and Random Forests
- Boosting
- Margins and over-fitting

Assignment Project Exam Help

https://eduassistpro.github.i

Add WeChat edu_assist_pr

Motivation

- Recall model averaging: given T real-valued predictors $\hat{f}^{(1)}, \ldots, \hat{f}^{(T)}$, form ensemble predictor $\hat{f}_{\text{avg}}$

Assignment Project Exam Help

$$\hat{f}_{\text{avg}}(x) := \frac{1}{T} \sum_{t=1}^{T} \hat{f}^{(t)}(x).$$

- https://eduassistpro.github.i

$$\text{mse}(f_{\text{avg}}) = \frac{1}{T} \sum_{t=1}^{T} \text{mse}(f^{t}) - \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[(\hat{f}_{\text{avg}}(X) - \hat{f}^{(t)}(X))^2\right].$$

Add WeChat edu_assist_pr

- For classification, analogue is *maj*

$$\hat{f}_{\text{maj}}(x) := \begin{cases} +1 & \text{if } \sum_{t=1}^{T} \hat{f}^{(t)}(x) > 0 \\ -1 & \text{otherwise} \end{cases}$$

($\hat{f}_{\text{avg}}$ is the scoring function used for $\hat{f}_{\text{maj}}$)

How to get classifiers to combine?

- Starting anew; how should we train classifiers to combine in majority-vote?

Assignment Project Exam Help

https://eduassistpro.github.i

Add WeChat edu_assist_pr

How to get classifiers to combine?

- Starting anew; how should we train classifiers to combine in majority-vote?
- Recall: model averaging works well when $\hat{}(t)$

Assignment Project Exam Help

https://eduassistpro.github.i

Add WeChat edu_assist_pr

How to get classifiers to combine?

- Starting anew; how should we train classifiers to combine in majority-vote?
- Recall: model averaging works well when $\hat{}(t)$
- To second point, learning algorithm shou variance"

Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

Using the same learning algorithm multiple times I

- Running same learning algorithm T times on the same data set yields T identical classifiers – not helpful!
- Instead, want to run same learning algorithm on T separate data sets.

Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

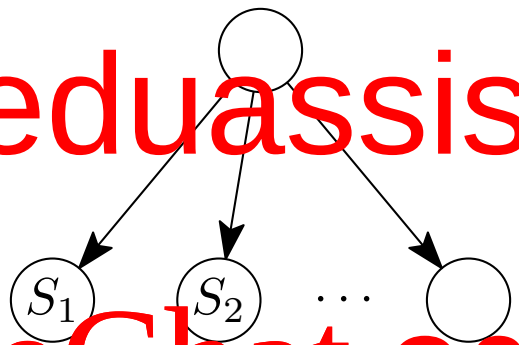


Figure 1: What we want is T da

Using the same learning algorithm multiple times II

- Invoke plug-in principle
  - In IID model, regard empirical disitribution on training examples P_n as estimate of the example distribution P.
  - Draw T independent data sets from P_n, and run learning algorithm on each data set.

Assignment Project Exam Help

https://eduassistpro.github.i

Add WeChat edu_assist_pr

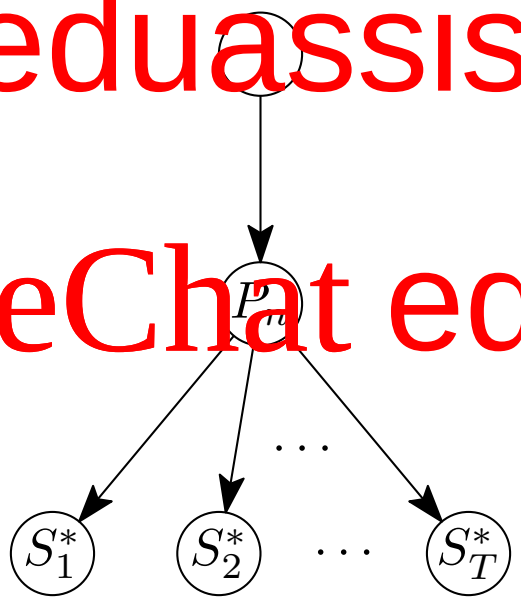


Figure 2: What we can get is T data sets from P_n

Bagging

- *Bagging*: bootstrap aggregating (Breiman, 1994)
- Given training data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \{-1, +1\}$
- For $t = 1, \ldots, T$:
  - Randomly draw n examples with replacement from training

$$\underset{(t)}{} \quad \underset{(t)}{} \quad \underset{n}{}$$

Assignment Project Exam Help

https://eduassistpro.github.i

Add WeChat edu_assist_pr

Aside: Sampling with replacement

- Pick n individuals from a population of size n with replacement.
- What is the chance that a given individual is not picked?
-

Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

trained on bootstrap sample

Random forests

- *Random Forests* (Breiman, 2001): Bagging with randomized variant of decision tree learning algorithm
  - Each time we need to choose a split, pick random subset of $\sqrt{d}$ features and only choose split from among those features.
-

Assignment Project Exam Help

https://eduassistpro.github.i

Add WeChat edu_assist_pr

Classifiers with independent errors

- Say we have T binary classifiers $\hat{f}^{(1)}, \ldots, \hat{f}^{(T)}$
- Assume on a given x, each provides an incorrect prediction with probability 0.4:

Assignment Project Exam Help

$$\Pr(\hat{f}^{(t)}(X) \neq Y \quad X = x) = 0.4.$$

https://eduassistpro.github.i

- What is chance that more than half of the class incorrect prediction?

Add WeChat edu_assist_pr

Coping with non-independent errors

- Classifier errors are unlikely to be independent; do something else to benefit from majority-vote

Assignment Project Exam Help

https://eduassistpro.github.i

Add WeChat edu_assist_pr

Coping with non-independent errors

- Classifier errors are unlikely to be independent; do something else to benefit from majority-vote
- Change how we obtain the individual classifiers:

Assignment Project Exam Help

https://eduassistpro.github.i

Add WeChat edu_assist_pr

Coping with non-independent errors

- Classifier errors are unlikely to be independent; do something else to benefit from majority-vote
- Change how we obtain the individual classifiers:
- Loop:
  - Use learning algorithm to get new classifi
  - Re-weight training examples to emph new classifier is incorrect

Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

Adaptive Boosting

- *AdaBoost* (Freund and Schapire, 1997)
- Training data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \{-1, +1\}$
- Initialize $D_1(i) = 1/n$ for all $i = 1, \ldots, n$
- For $t = 1, \ldots, T$:
  - Run learning algorithm on D_t-weighted training examples, get

$$z_t \qquad D_t\ i\ \cdot y_i f \quad x_i \qquad ,$$

$$\alpha := \frac{1}{2} \ln \frac{1 + z_t}{1 - z_t} \in \mathbb{R}$$

$$D_{t+1}(i) := \frac{D_t(i) \exp(-\alpha_t \cdot y_i f^{\ t}\ \ x_i}{Z_t} \quad \text{for } i = 1, \ldots, n.$$

Here Z_t is normalizer that makes D_{t+1} a probability distribution.

- Final classifier: $\hat{f}(x) = \text{sign}(\sum_{t=1}^{T} \alpha_t \cdot f^{(t)}(x))$

Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

Plot of α_t as function of z_t

$$\alpha_t = \frac{1}{2} \ln \frac{1 + z_t}{1 - z_t} \in \mathbb{R}$$

Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

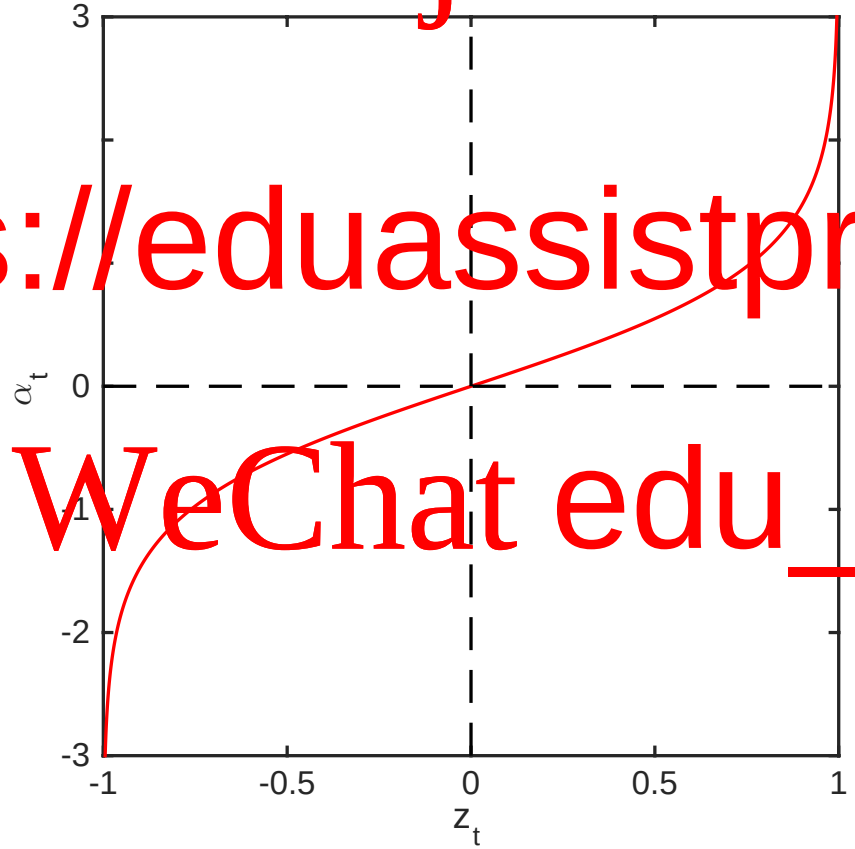


Figure 3: α_t as function of z_t

Example: AdaBoost with decision stumps

- (From Figures 1.1 and 2.2 of Schapire & Freund text.)
- Use "decision stump" learning algorithm with AdaBoost
  - Each $f^{(t)}$ has the form

$$f^{(t)}(x) = \begin{cases} +1 & \text{if } x_i > \theta \\ -1 & \text{if } x_i \leq \theta \end{cases} \quad \text{or} \quad f^{(t)}(x) = \begin{cases} -1 & \text{if } x_i > \theta \\ +1 & \text{if } x_i \leq \theta \end{cases}$$

decision

Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

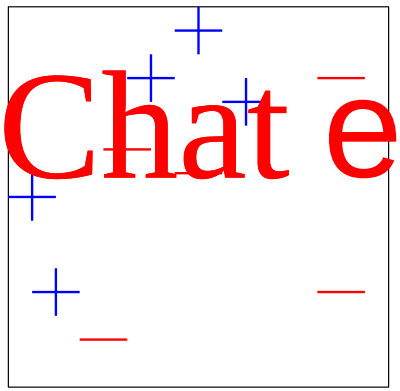
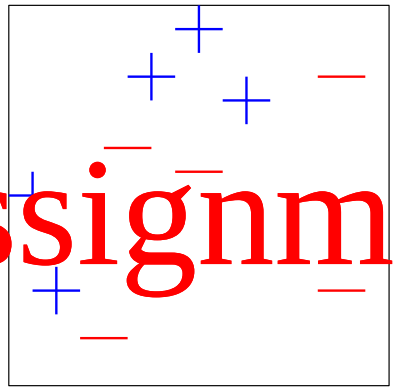


Figure 4: Training data for example execution

Example execution of AdaBoost I

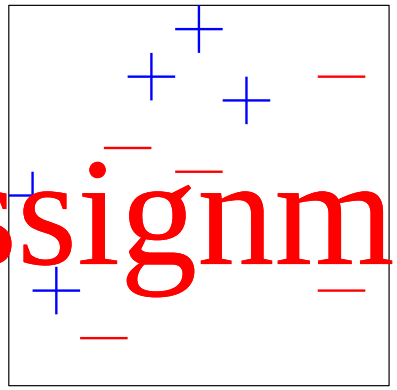


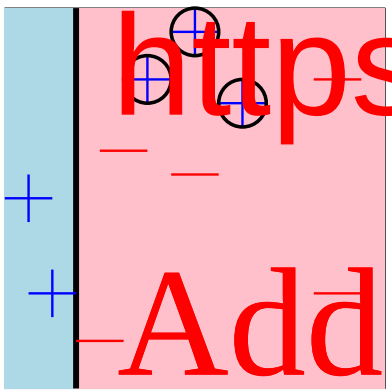
Assignment Project Exam Help

https://eduassistpro.github.i

Add WeChat edu_assist_pr

Example execution of AdaBoost II



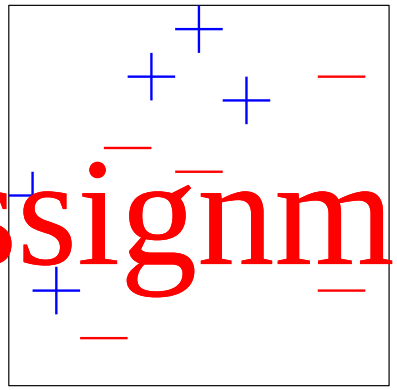


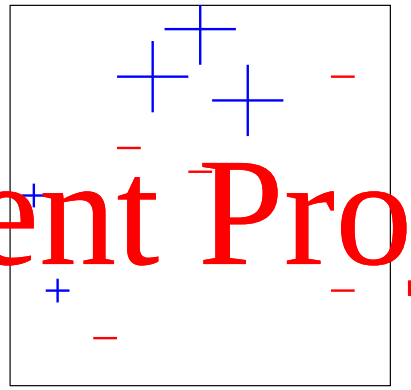
$f^{(1)}$

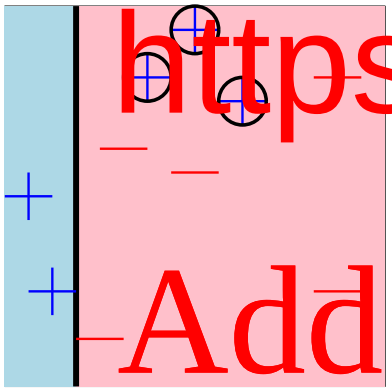
$z_1 = 0.40,\ \alpha_1 = 0.42$

Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

Example execution of AdaBoost III





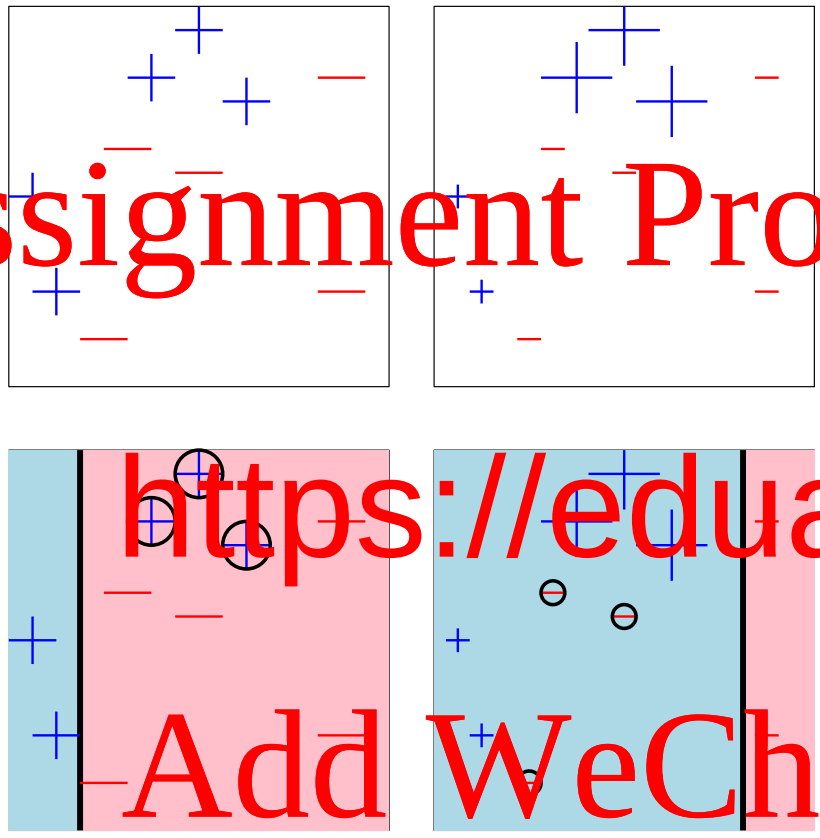


$f^{(1)}$

$z_1 = 0.40,\ \alpha_1 = 0.42$

Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

Example execution of AdaBoost IV

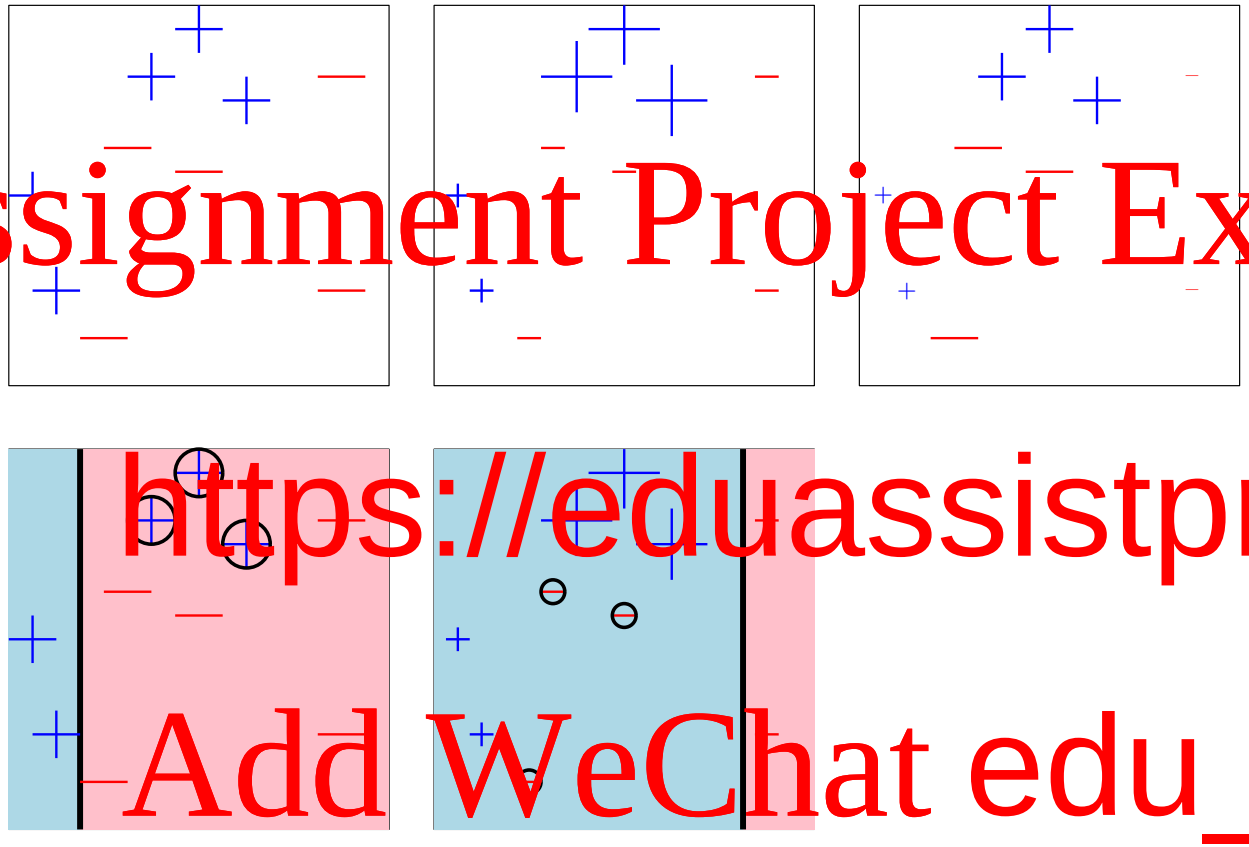


$f^{(1)}$
$z_1 = 0.40,\ \alpha_1 = 0.42$

$f^{(2)}$
$z_2 = 0.58,\ \alpha_2 = 0.65$

Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

Example execution of AdaBoost V

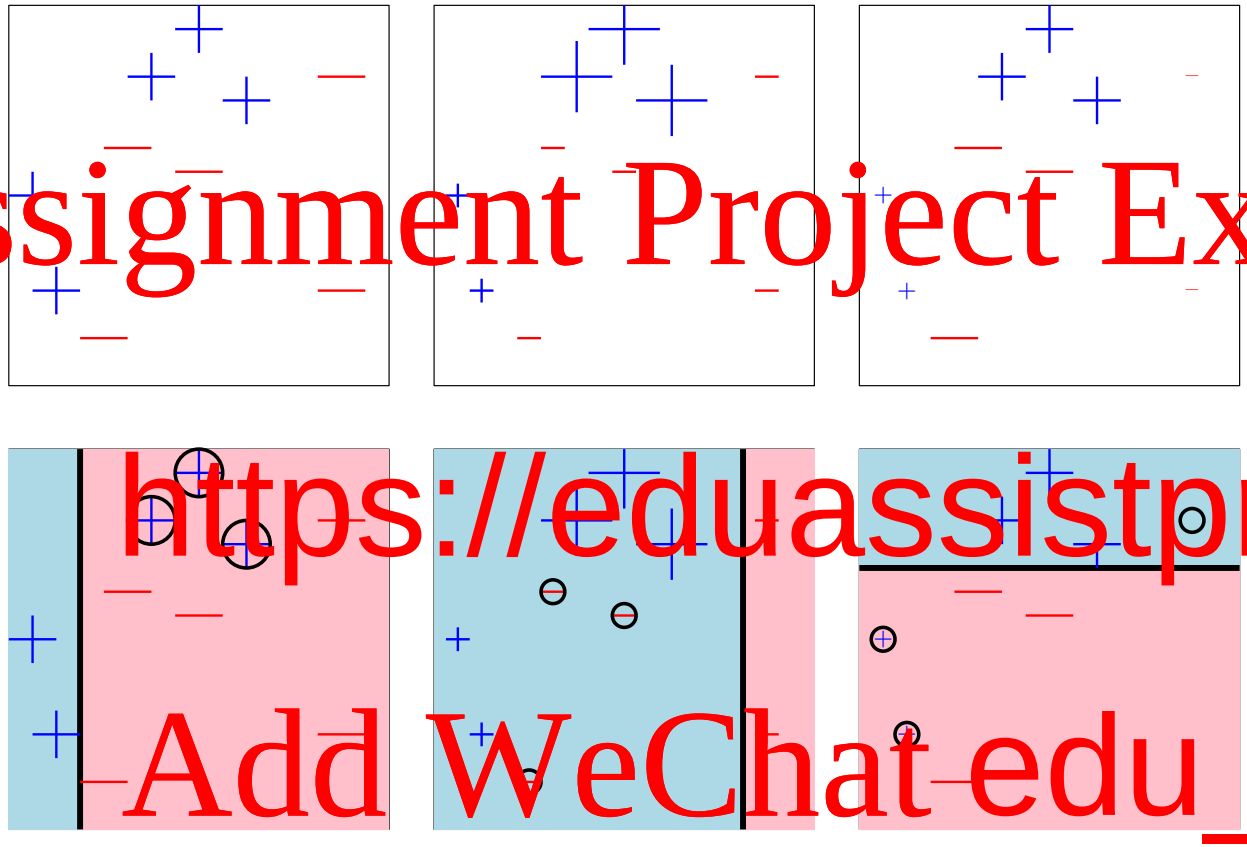


Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

$f^{(1)}$
$z_1 = 0.40,\ \alpha_1 = 0.42$

$f^{(2)}$
$z_2 = 0.58,\ \alpha_2 = 0.65$

Example execution of AdaBoost VI



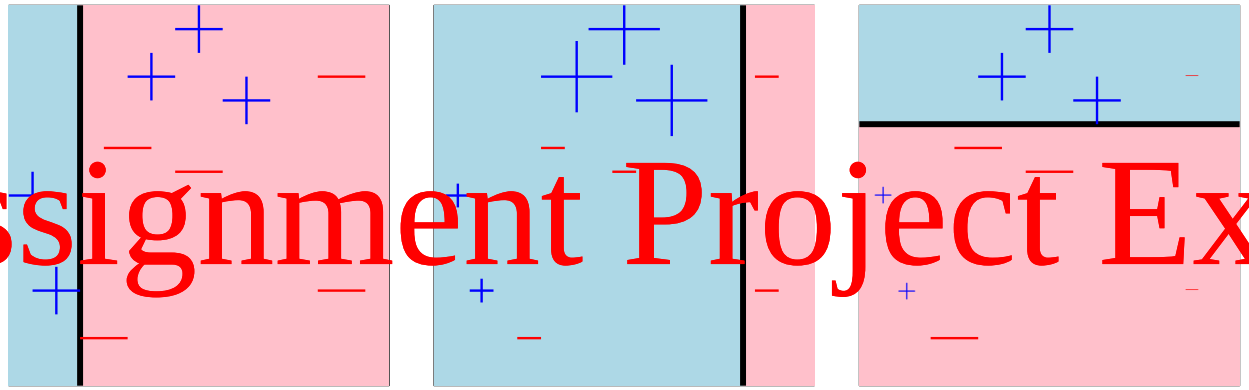
Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

$f^{(1)}$
$z_1 = 0.40,\ \alpha_1 = 0.42$

$f^{(2)}$
$z_2 = 0.58,\ \alpha_2 = 0.65$

$f^{(3)}$
$z_3 = 0.72,\ \alpha_3 = 0.92$

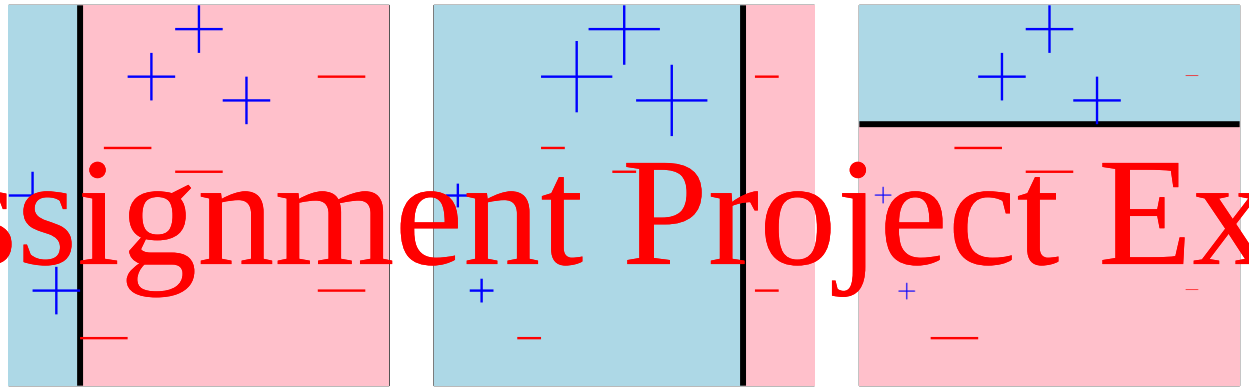
Example execution of AdaBoost VII



$z_1 = 0$

Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

Example execution of AdaBoost VIII



Assignment Project Exam Help

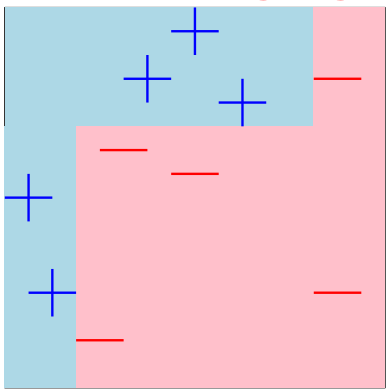
https://eduassistpro.github.i

$z_1 = 0$

Final classifier:

$\hat{f}(x) = \text{sign}\left(0.42 f^{(1)}(x) + 0.65 f^{(2)}(x\right.$

Add WeChat edu_assist_pr



Training error rate of final classifier

- Let $\gamma_t := z_t/2$: advantage over random guessing achieved by $f^{(t)}$
- **Theorem.** Training error rate of final classifier is

$$\leq \exp\left(-2\sum_{t=1}^{T}\gamma_t^2\right) \leq \exp\left(-2\bar{\gamma}^2 T\right)$$

where

$$\bar{\gamma}^2 := \frac{1}{T}\sum_{t=1}^{T}\gamma_t^2$$

Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

Training error rate of final classifier

- Let $\gamma_t := z_t/2$: advantage over random guessing achieved by $f^{(t)}$
- **Theorem.** Training error rate of final classifier is

$$\leq \exp\left(-2\sum_{t=1}^{T}\gamma_t^2\right) = \exp\left(-2\bar{\gamma}^2 T\right)$$

where

$$\bar{\gamma}^2 := \frac{1}{T}\sum_{t=1}^{T}\gamma_t^2$$

- AdaBoost is "adaptive":
  - Some γ_t can be small (even negative)—only care about average $\bar{\gamma}^2$

Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

Training error rate of final classifier

- Let $\gamma_t := z_t/2$: advantage over random guessing achieved by $f^{(t)}$
- **Theorem.** Training error rate of final classifier is

$$\leq \exp\left(-2\sum_{t=1}^{T}\gamma_t^2\right) = \exp\left(-2\bar{\gamma}^2 T\right)$$

where

$$\bar{\gamma}^2 := \frac{1}{T}\sum_{t=1}^{T}\gamma_t^2$$

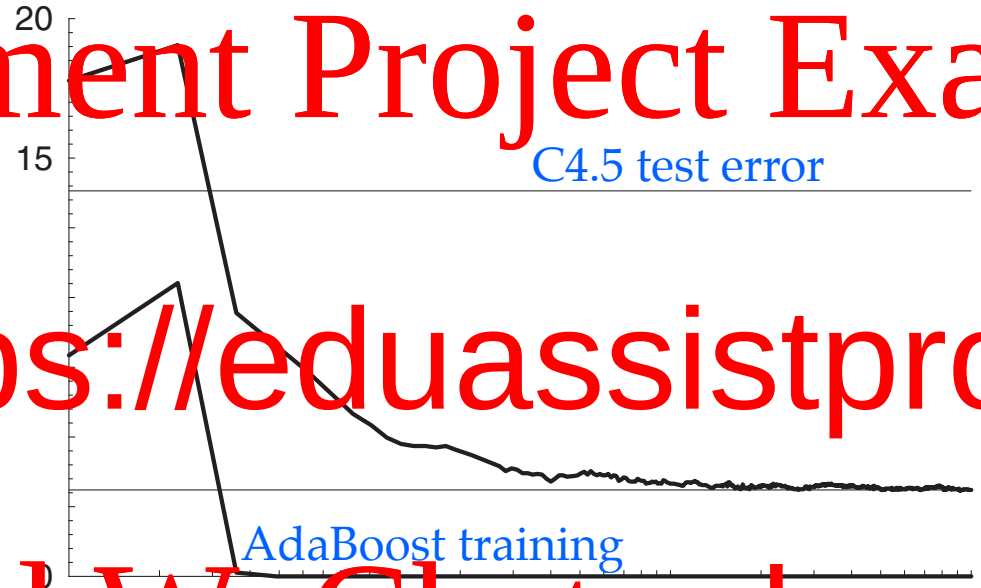
- AdaBoost is "adaptive":
  - Some γ_t can be small (even negative)—only care about average $\bar{\gamma}^2$
- What about true error rate in IID model?
  - A very complex model as T becomes large!

Assignment Project Exam Help
https://eduassistpro.github.i
Add WeChat edu_assist_pr

Surprising behavior of boosting

- AdaBoost + C4.5 decision tree learning on "letters" data set

Assignment Project Exam Help



https://eduassistpro.github.i
Add WeChat edu_assist_pr

Figure 5: Figure 1.7 from Schapire & Fre

- Training error rate is zero after five iterations.
- Test error rate continues to decrease, even up to 1000 iterations.

Margins theory

- Look at (normalized) scoring function of final classifier

$$\hat{h}(x) = \frac{\sum_{t=1}^{T} \alpha_t \cdot f^{(t)}(x)}{\sum_{t=1}^{T} |\alpha_t|} \in [-1, +1].$$

- Say $y \ \ \hat{h}(x)$ is margin achieved by $\hat{h}$ on example (x, y)

Assignment Project Exam Help

https://eduassistpro.github.i

Add WeChat edu_assist_pr

Margins theory

- Look at (normalized) scoring function of final classifier

$$\hat{h}(x) = \frac{\sum_{t=1}^{T} \alpha_t \cdot f^{(t)}(x)}{\sum_{t=1}^{T} |\alpha_t|} \in [-1, +1].$$

Assignment Project Exam Help

- Say $y \quad \hat{h}(x)$ is margin achieved by $\hat{h}$ on example (x, y)
- https://eduassistpro.github.i
  - AdaBoost tends to increase margins on t
- (Similar to but not same as SVM margins)

Add WeChat edu_assist_pr

Margins theory

- Look at (normalized) scoring function of final classifier

$$\hat{h}(x) = \frac{\sum_{t=1}^{T} \alpha_t \cdot f^{(t)}(x)}{\sum_{t=1}^{T} |\alpha_t|} \in [-1, +1].$$

Assignment Project Exam Help

- Say $y \quad \hat{h}(x)$ is margin achieved by $\hat{h}$ on example (x, y)
- https://eduassistpro.github.i
  - AdaBoost tends to increase margins on t
- (Similar to but not same as SVM margins)
- Add WeChat edu_assist_pr
- On "letters" data set

	$T = 5$	$T = 100$	$T = 1000$
training error rate	0.0%	0.0%	0.0%
test error rate	8.4%	3.3%	3.1%
% margins ≤ 0.5	7.7%	0.0%	0.0%
min margin achieved	0.14	0.52	0.55