Data Weighting and Selection

Ling Li, Hsuan-Tien Lin, Amrit Pratap, David Soloveichik, Yaser S. Abu-Mostafa

Learning Systems Group

http://www.work.caltech.edu

ρ Learning

If an example tends to contradict the hypotheses that the learning process can produce, it can be harmful to learning. Let f be the target function, $e(g(x), f(x))$ be the 0/1 loss function, and $\pi(g)$ be the out-of-sample error. For a particular example x, how well $e(g(x), f(x))$ correlates to $\pi(g)$ can be a measure of how "good" the example is. We define

$$\rho(x) = \mathrm{corr}_g[e(g(x), f(x)), \pi(g)].$$

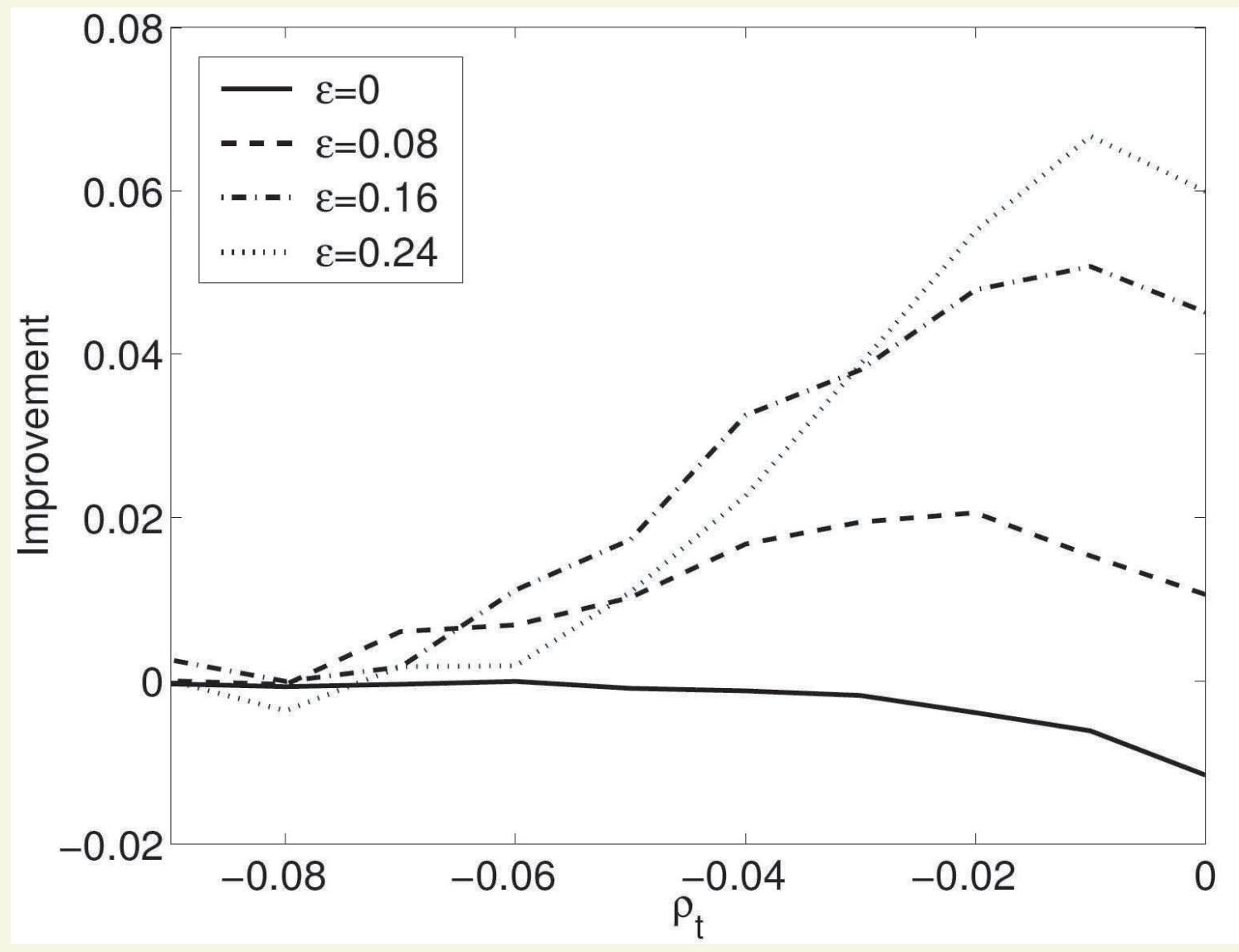
In a symmetric learning model,

$$\rho(x) \propto E_g[\pi(g) | g(x) \neq f(x)] - E_g[\pi(g) | g(x) = f(x)].$$

Intuitively, examples have a high ρ value if getting them right helps to get a smaller $\pi(g)$. We want to reweigh examples based on ρ, assigning higher weights to the more representative examples with high ρ value.

It is easy to show that for a data sample of size 1, weighting examples proportionally to $1 - E_g[\pi(g) | g(x) = f(x)]$ provably decreases the expected out-of-sample error. Extending the proof to larger data samples requires that the out-of-sample performance of hypotheses that do well on a single example in T is correlated, on average, to the performance of those that do well on multiple examples in T.

In practice, f is not known. However, π can be estimated based on the leave-one-out error estimate. This estimation seems to work well in the case of noise, and can be used to do data selection successfully.



Generalization error improvement using ρ based data selection. Examples with $\rho < \rho_t$ are discarded. ε is the noise level. [From: A. Nicholson 02]

Boosting

Boosting is a general framework to improve the accuracy of any "weak" learner by iteratively generating a linear combination of weak hypotheses. It maintains a set of weights over the training set and emphasizes hard examples by giving them higher weights and favors hypotheses with lower training error by giving them larger coefficients in the linear combination.

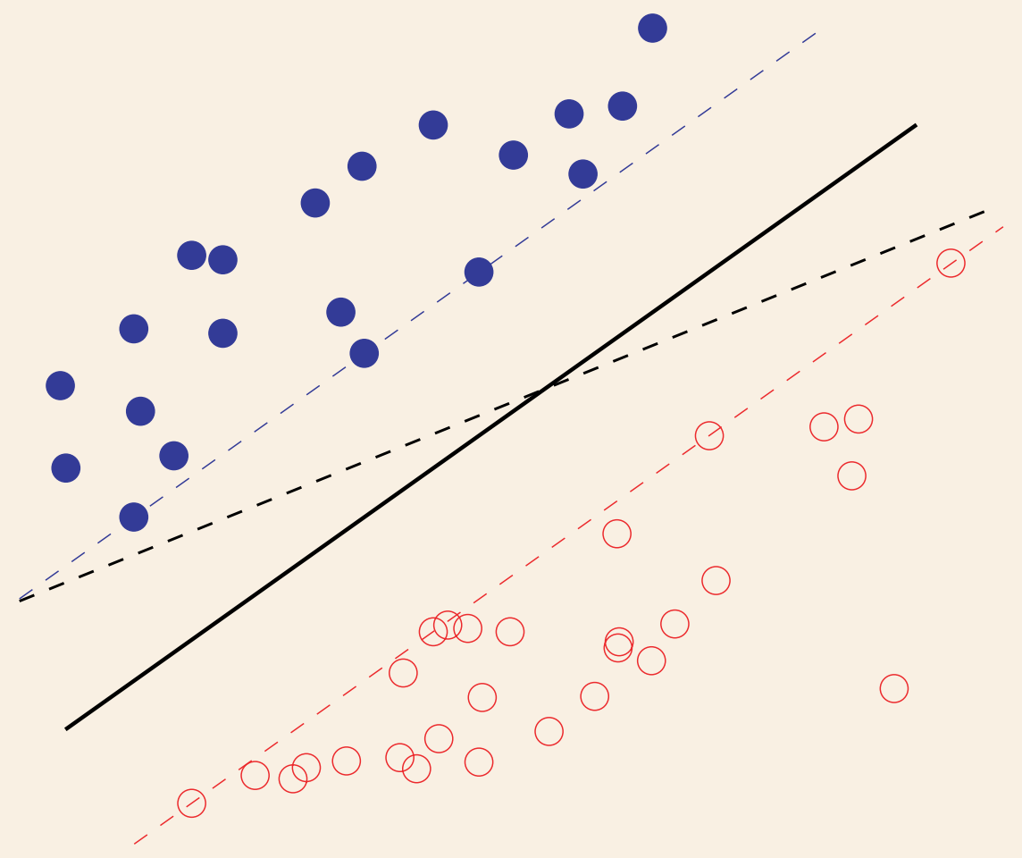
```
• Given $S = \{(x_i, y_i)\}_{i=1}^{n}$. Initialize $w_i^1 = 1/n$.
• For $t = 1$ to $T$:
    – Train weak learner with weights $w_i^t \rightarrow h_t$
    – Set $\alpha_t$ based on the performance of $h_t$
    – Set $w_i^{t+1} \propto w_i^t e^{-\alpha_t y_i h_t(x_i)}$
• Output sign $\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$ as the combined hypothesis.
```

The sequence of weights $w_i^t \propto \prod_{j=1}^{t} e^{-\alpha_j y_i h_j(x_i)}$ can be used as a measure of how "hard" it was to get an example right. Examples that are too hard to get might be outliers or noise. The average weight is used as a measure of hardness.

Support Vector Machines

SVM is a learning algorithm that finds a linear function with largest margin to separate data. The problem is generally solved in Lagrange dual space:

$$\min_{\alpha} \frac{1}{2}\alpha^T Q \alpha + e^T \alpha$$
$$y^T \alpha = 0, 0 \leq \alpha \leq C$$
$$Q_{ij} \equiv y_i y_j \langle x_i, x_j \rangle$$

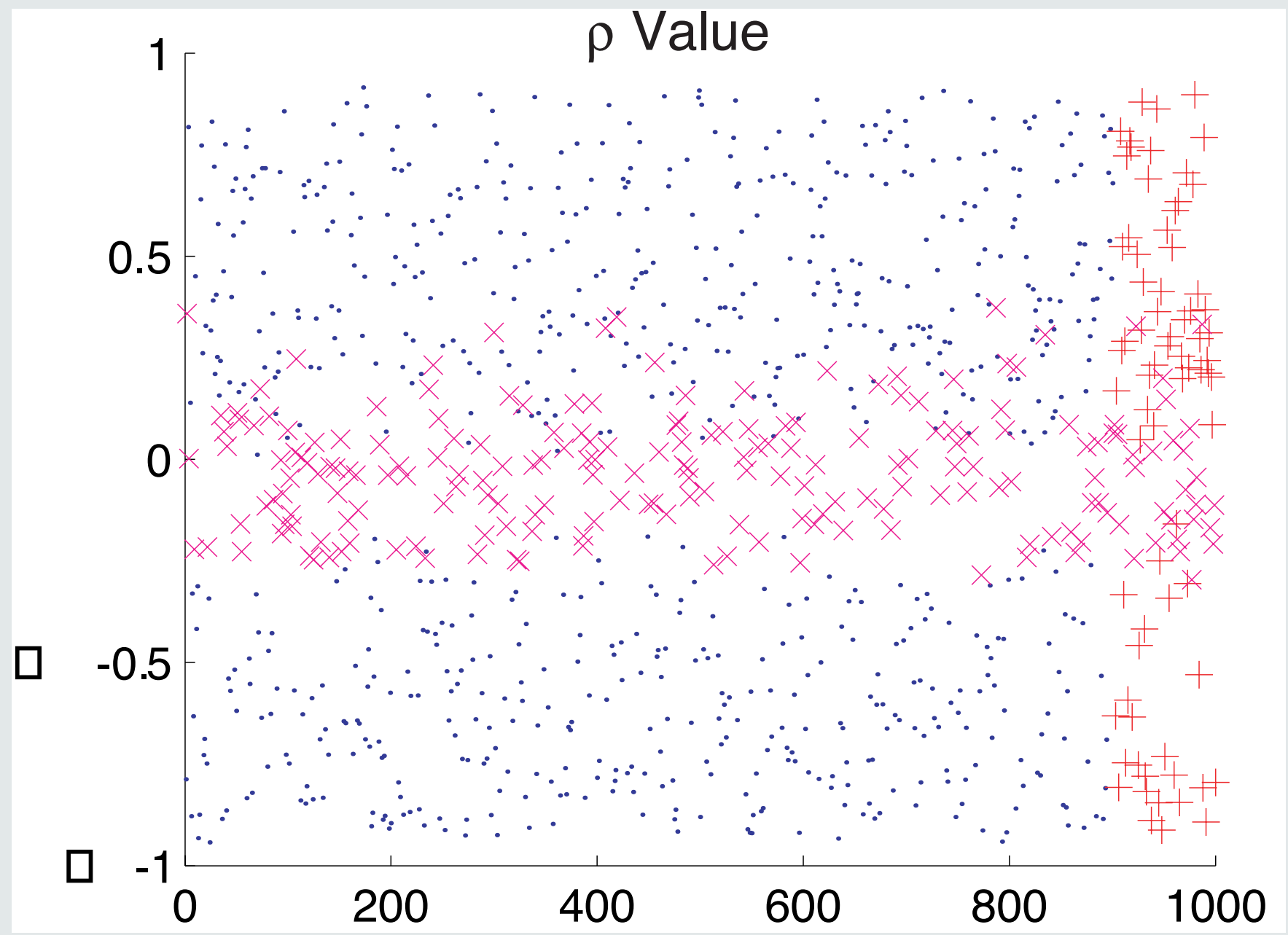


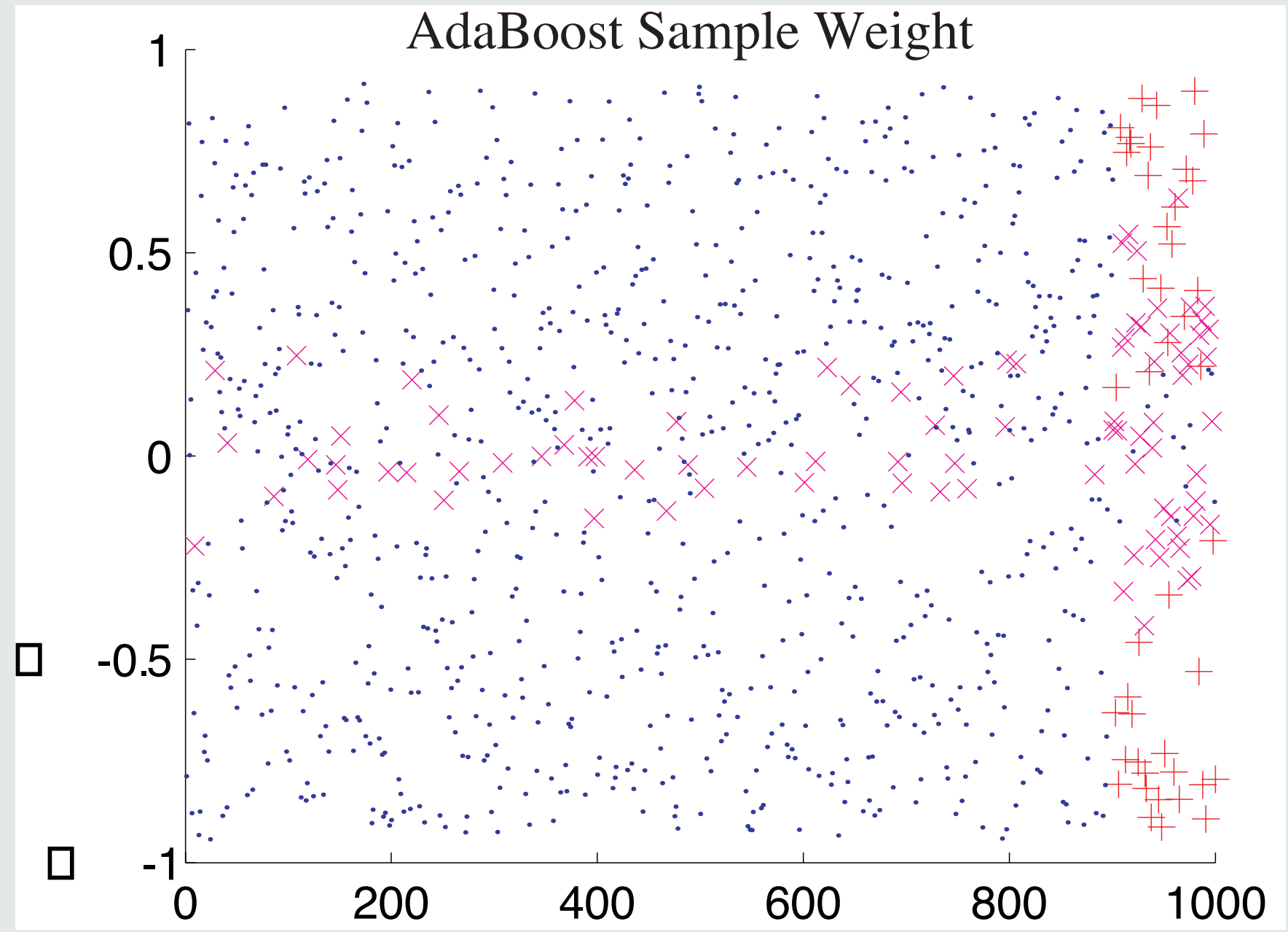
The SVM decision function is of the form $\sum \alpha_i y_i \langle x_i, x \rangle + b$, so the Lagrange multiplier α_i indicates whether the data x_i is separated correctly or not. That is, the SVM decision function would give zero weights to "trivially separable" examples, and maximum weights to "violating" examples. The degree of hardness of an example can be calculated from its violation of the margin.

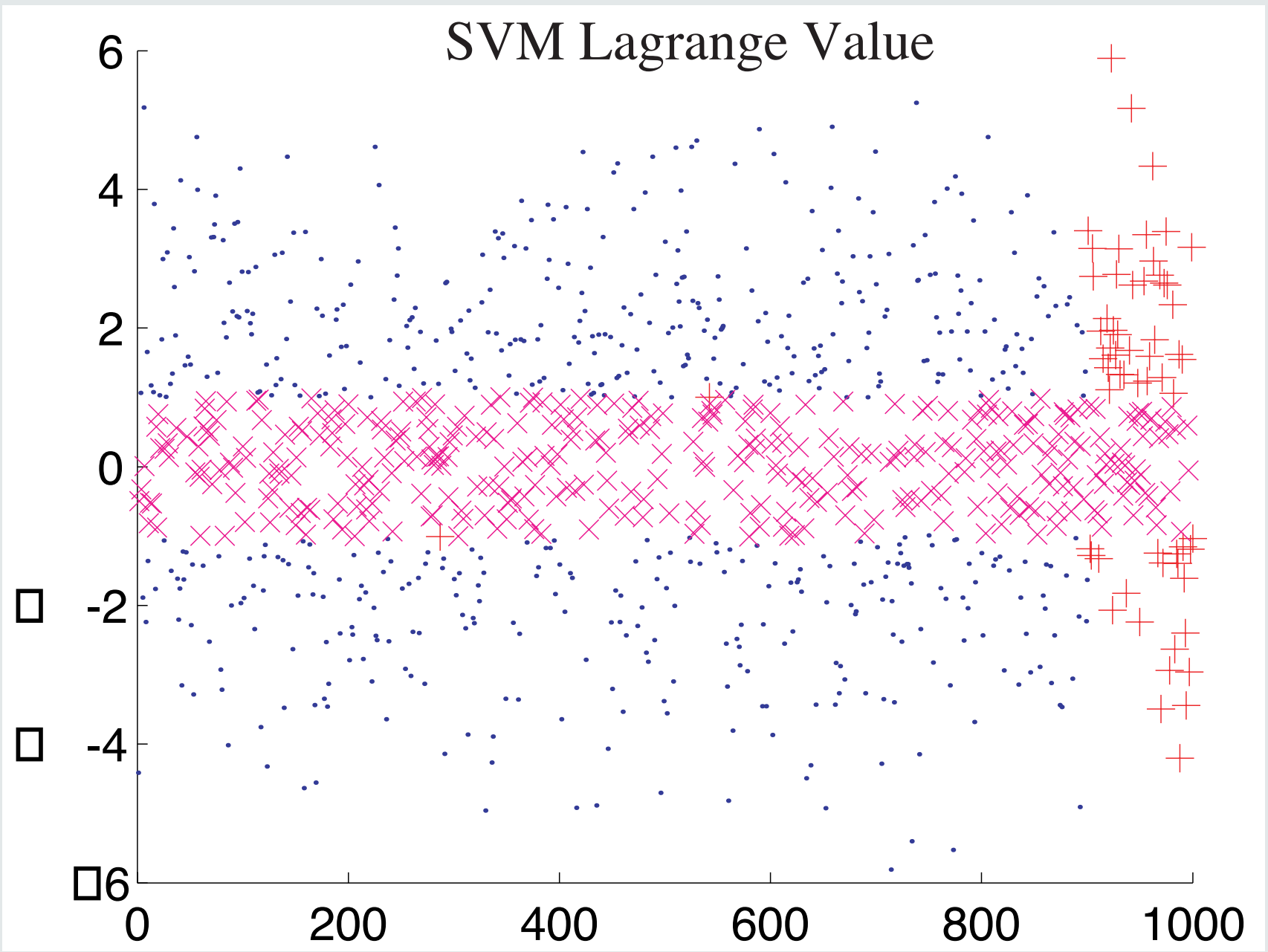
Fingerprinting Experiments

We designed a fingerprint experiment to identify characteristics of examples that relate to learning. We randomly generated training examples and "flipped" the last 10% of them. Three approaches (ρ values, AdaBoost sample weights, and SVM Lagrange values) rooted from different parts of machine learning essentially gave the same results.

That is, examples can be divided into roughly three categories: normal (blue ·), difficult (magenta ×), and noisy (red +). Noisy examples are usually bad and difficult examples are usually boundary cases to learning. Properly weighting these categories of examples may benefit the learning process.


ρ Value


AdaBoost Sample Weight


SVM Lagrange Value