Paper 293: Online Algorithms for Multiclass Classification using Partial Labels

Rajarshi Bhattacharjee [1] and Naresh Manwani [2]

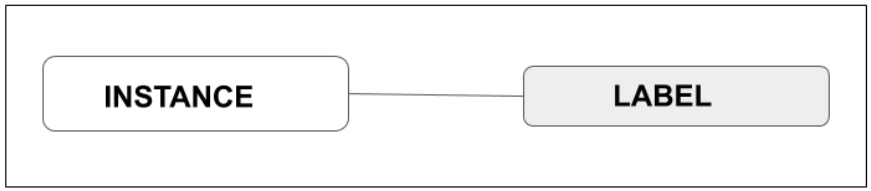
[1]IIT Madras

[2]IIIT Hyderabad

PAKDD 2020

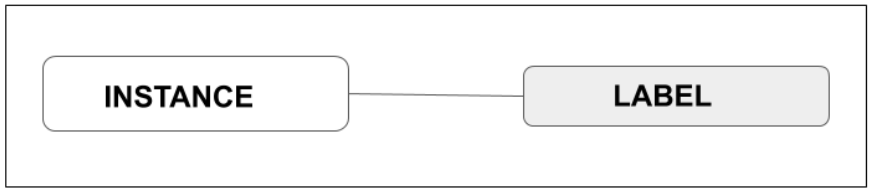
Overview

Strong vs Partial Supervised Learning

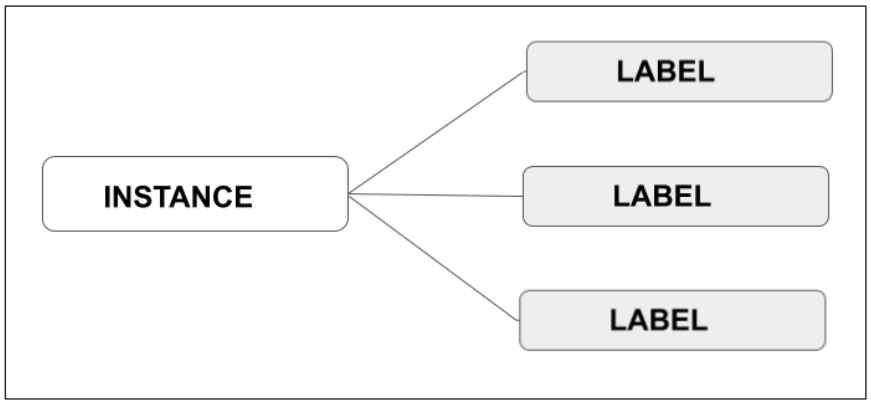


- **Strong Supervision**: **Single** label associated with an instance
- *Easy* task

Strong vs Partial Supervised Learning

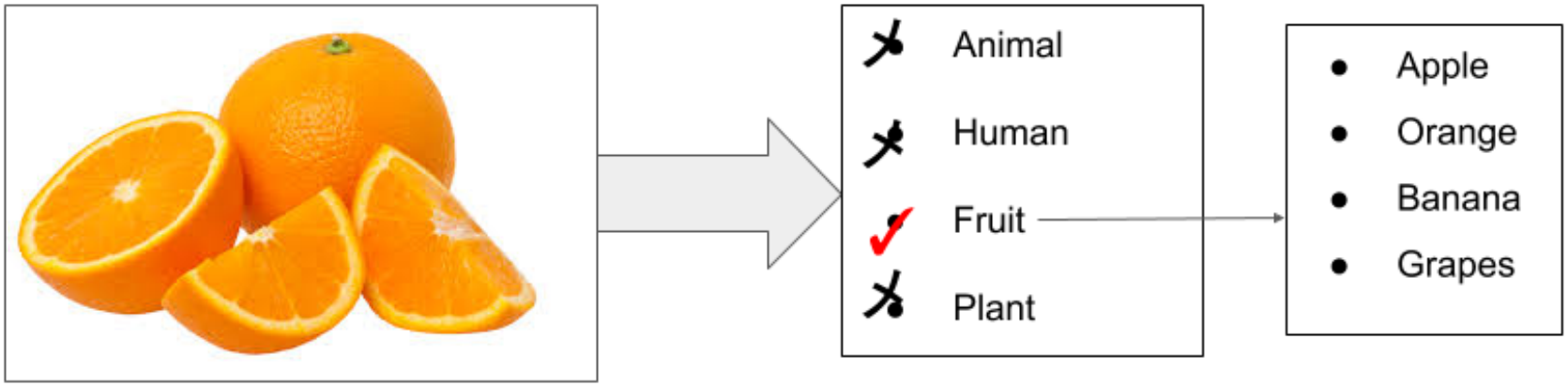


- **Strong Supervision**: **Single** label associated with an instance
- *Easy* task



- **Partial/Weak Supervision**: **Multiple** labels associated with an instance
- Only one label in the label set is the **true** label
- *Hard* task

Why Partial Supervision?



- Partial Labels are **cheap**
- Labeling less time-consuming
- Example: Automatic Image Annotation

Online Learning Algorithms

- Online Learning Algorithms process data *sequentially*

Online Learning Algorithms

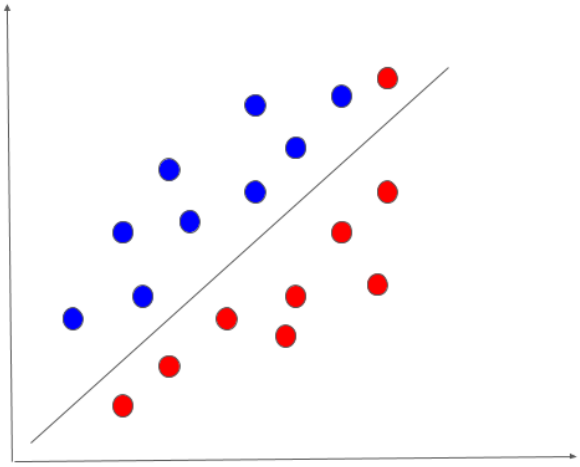
- Online Learning Algorithms process data *sequentially*
- Data can be generated by *non-stationary* distribution.

Online Learning Algorithms

- Online Learning Algorithms process data *sequentially*
- Data can be generated by *non-stationary* distribution.
- Useful in streaming settings with real-time data like online trading.

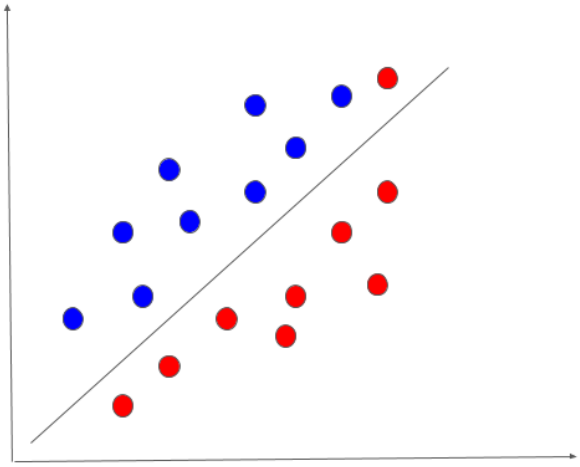
Online Learning Algorithms

- Online Learning Algorithms process data *sequentially*
- Data can be generated by *non-stationary* distribution.
- Useful in streaming settings with real-time data like online trading.
- Popular margin-based online classifiers include **Pegasos** and **Perceptron**.



Online Learning Algorithms

- Online Learning Algorithms process data *sequentially*
- Data can be generated by *non-stationary* distribution.
- Useful in streaming settings with real-time data like online trading.
- Popular margin-based online classifiers include **Pegasos** and **Perceptron**.
- We propose the first online algorithms for multiclass classification with Partial Labels based on Pegasos and Perceptron.



Problem Statement

- $\mathcal{X} \subseteq \mathbb{R}^d$: the feature space from which the instances are drawn
- $\mathcal{Y} = \{1, \ldots, K\}$: the output label space
- Every instance $\mathbf{x} \in \mathcal{X}$ is associated with a candidate label set $Y \subseteq \mathcal{Y}$
- $\overline{Y}$: The set of labels not present in the candidate label set ($Y \cup \overline{Y} = [K]$)

Problem Statement

- $\mathcal{X} \subseteq \mathbb{R}^d$: the feature space from which the instances are drawn
- $\mathcal{Y} = \{1, \ldots, K\}$: the output label space
- Every instance $\mathbf{x} \in \mathcal{X}$ is associated with a candidate label set $Y \subseteq \mathcal{Y}$
- $\overline{Y}$: The set of labels not present in the candidate label set ($Y \cup \overline{Y} = [K]$)
- **Objective**: Learn a *linear* classifier $h : \mathcal{X} \to \mathcal{Y}$ parameterized by weights $W \in \mathbb{R}^{d \times K}$ defined as:

$$h(\mathbf{x}) = \arg\max_{i \in [K]} \ \mathbf{w}_i.\mathbf{x}$$

where $\mathbf{w}_i$ (ith column vector of W) denotes the parameter vector corresponding to the i^{th} class.

Label Disambiguation for Partial Label Set

- Disambiguation by **Averaging**
  - All candidate labels are equally likely
  - Prediction based on averaging the output
  - ***Average Prediction* Hinge Loss**: A representative convex margin-based loss function ($[x]_+ = max(x, 0)$)

$$L_{APH}(h(\mathbf{x}), Y) = \left[1 - \frac{1}{|Y|}\sum_{i \in Y} \mathbf{w}_i.\mathbf{x} + \max_{j \notin Y} \mathbf{w}_j.\mathbf{x}\right]_+$$

Label Disambiguation for Partial Label Set

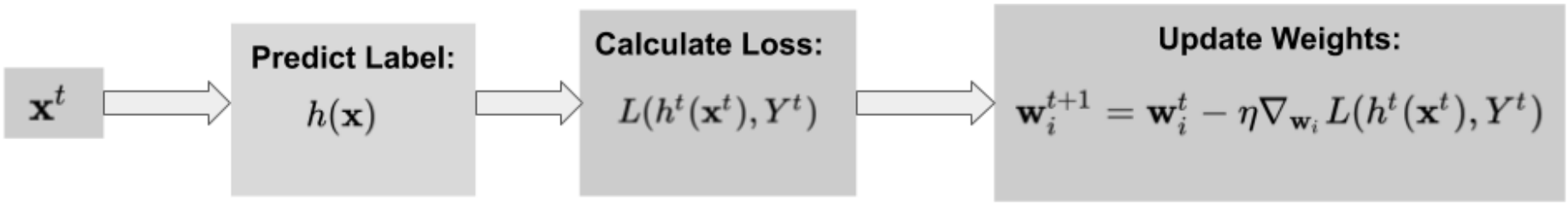
- Disambiguation by **Averaging**
  - All candidate labels are equally likely
  - Prediction based on averaging the output
  - ***Average Prediction* Hinge Loss**: A representative convex margin-based loss function ($[x]_+ = max(x, 0)$)

$$L_{APH}(h(\mathbf{x}), Y) = \left[1 - \frac{1}{|Y|}\sum_{i \in Y} \mathbf{w}_i.\mathbf{x} + \max_{j \notin Y} \mathbf{w}_j.\mathbf{x}\right]_+$$

- Disambiguation by **Identification**
  - Ground truth is a latent variable
  - Iterative refining to predict ground truth
  - ***Max Prediction* Hinge Loss (MPH)**: A representative non-convex margin-based loss function

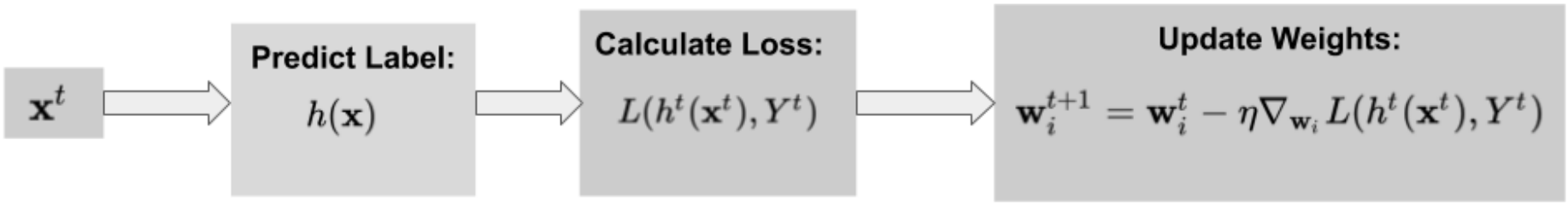
$$L_{MPH}(h(\mathbf{x}), Y) = \left[1 - \max_{i \in Y} \mathbf{w}_i.\mathbf{x} + \max_{j \notin Y} \mathbf{w}_j.\mathbf{x}\right]_+$$

Avg Perceptron and Max Perceptron



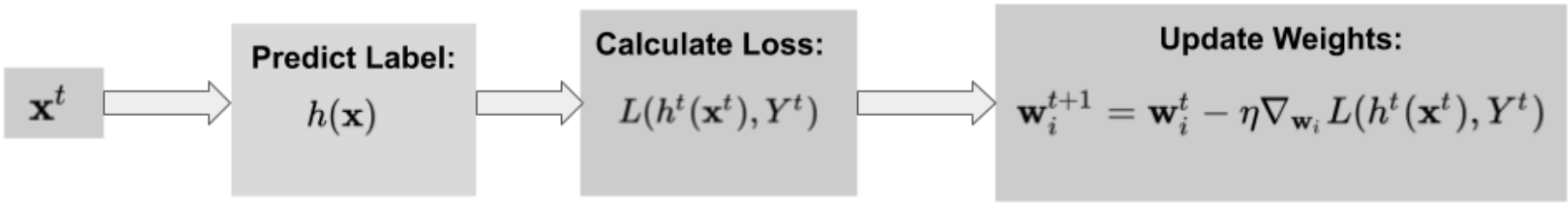
- Input at t: $\mathbf{x}^t$

Avg Perceptron and Max Perceptron



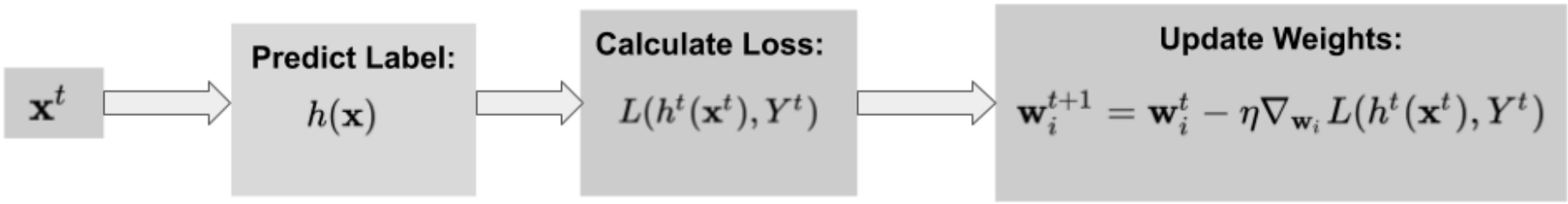
- Input at t: $\mathbf{x}^t$
- Partial label set Y^t is revealed *after* label is predicted

Avg Perceptron and Max Perceptron



- Input at t: $\mathbf{x}^t$
- Partial label set Y^t is revealed *after* label is predicted
- Weights are updated using **sub-gradients** of appropriate Loss function

Avg Perceptron and Max Perceptron



- Input at t: $\mathbf{x}^t$
- Partial label set Y^t is revealed *after* label is predicted
- Weights are updated using **sub-gradients** of appropriate Loss function
- Sub-gradients of the Avg and Max Prediction Loss are needed for the update rule.

Avg Perceptron Mistake Bound under Linear Separability

Theorem

Let $(\mathbf{x}^1, Y^1), \ldots, (\mathbf{x}^T, Y^T)$ *be the examples presented to Avg Perceptron, where* $\mathbf{x}^t \in \mathbb{R}^d$ *and* $Y^t \subseteq [K]$*. Let* $W^* \in \mathbb{R}^{d \times K}$ $(\|W^*\| = 1)$ *be such that* $\frac{1}{|Y^t|} \sum_{i \in Y^t} \mathbf{w}_i^* . \mathbf{x}^t - \max_{j \in \overline{Y}^t} \mathbf{w}_j^* . \mathbf{x}^t \geq \gamma$, $\forall t \in [T]$*. Then we have:*

$$\sum_{t=1}^{T} L_A(h^t(\mathbf{x}^t), Y^t) \leq \frac{2}{\gamma^2} + \left[\frac{1}{c} + 1\right] \frac{R^2}{\gamma^2}$$

where $L_A(h(\mathbf{x}), Y) = \mathbb{I}_{\{h(\mathbf{x}) \notin Y\}}$, $c = \min_t |Y^t|$, $R = \max_t ||\mathbf{x}^t||$ *and* $\gamma \geq 0$ *is the margin of separation.*

- L_A: The 0-1 partial loss c: minimum label set size γ:margin
- The bound is inversely proportional to the minimum label set size

Avg Perceptron Mistake Bound under Linear Separability

Theorem

Let $(\mathbf{x}^1, Y^1), \ldots, (\mathbf{x}^T, Y^T)$ *be the examples presented to Avg Perceptron, where* $\mathbf{x}^t \in \mathbb{R}^d$ *and* $Y^t \subseteq [K]$. *Let* $W^* \in \mathbb{R}^{d \times K}$ $(\|W^*\| = 1)$ *be such that* $\frac{1}{|Y^t|} \sum_{i \in Y^t} \mathbf{w}_i^* . \mathbf{x}^t - \max_{j \in \overline{Y}^t} \mathbf{w}_j^* . \mathbf{x}^t \geq \gamma$, $\forall t \in [T]$. *Then we have:*

$$\sum_{t=1}^{T} L_A(h^t(\mathbf{x}^t), Y^t) \leq \frac{2}{\gamma^2} + \left[\frac{1}{c} + 1\right] \frac{R^2}{\gamma^2}$$

where $L_A(h(\mathbf{x}), Y) = \mathbb{I}_{\{h(\mathbf{x}) \notin Y\}}$, $c = \min_t |Y^t|$, $R = \max_t ||\mathbf{x}^t||$ *and* $\gamma \geq 0$ *is the margin of separation.*

- L_A: The 0-1 partial loss c: minimum label set size γ:margin
- The bound is inversely proportional to the minimum label set size

Avg Perceptron Mistake Bound under Linear Separability

Theorem

Let $(\mathbf{x}^1, Y^1), \ldots, (\mathbf{x}^T, Y^T)$ *be the examples presented to Avg Perceptron, where* $\mathbf{x}^t \in \mathbb{R}^d$ *and* $Y^t \subseteq [K]$. *Let* $W^* \in \mathbb{R}^{d \times K}$ $(\|W^*\| = 1)$ *be such that* $\frac{1}{|Y^t|} \sum_{i \in Y^t} \mathbf{w}_i^* . \mathbf{x}^t - \max_{j \in \overline{Y}^t} \mathbf{w}_j^* . \mathbf{x}^t \geq \gamma, \ \forall t \in [T]$. *Then we have:*

$$\sum_{t=1}^{T} L_A(h^t(\mathbf{x}^t), Y^t) \leq \frac{2}{\gamma^2} + \left[\frac{1}{c} + 1\right] \frac{R^2}{\gamma^2}$$

where $L_A(h(\mathbf{x}), Y) = \mathbb{I}_{\{h(\mathbf{x}) \notin Y\}}$, $c = \min_t |Y^t|$, $R = \max_t ||\mathbf{x}^t||$ *and* $\gamma \geq 0$ *is the margin of separation.*

- L_A: The 0-1 partial loss c: minimum label set size γ:margin
- The bound is inversely proportional to the minimum label set size

Avg Perceptron Mistake Bound under Linear Separability

Theorem

Let $(\mathbf{x}^1, Y^1), \ldots, (\mathbf{x}^T, Y^T)$ *be the examples presented to Avg Perceptron, where* $\mathbf{x}^t \in \mathbb{R}^d$ *and* $Y^t \subseteq [K]$. *Let* $W^* \in \mathbb{R}^{d \times K}$ $(\|W^*\| = 1)$ *be such that* $\frac{1}{|Y^t|} \sum_{i \in Y^t} \mathbf{w}_i^* . \mathbf{x}^t - \max_{j \in \overline{Y}^t} \mathbf{w}_j^* . \mathbf{x}^t \geq \gamma$, $\forall t \in [T]$. *Then we have:*

$$\sum_{t=1}^{T} L_A(h^t(\mathbf{x}^t), Y^t) \leq \frac{2}{\gamma^2} + \left[\frac{1}{c} + 1\right] \frac{R^2}{\gamma^2}$$

where $L_A(h(\mathbf{x}), Y) = \mathbb{I}_{\{h(\mathbf{x}) \notin Y\}}$, $c = \min_t |Y^t|$, $R = \max_t ||\mathbf{x}^t||$ *and* $\gamma \geq 0$ *is the margin of separation.*

- L_A: The 0-1 partial loss c: minimum label set size γ:margin
- The bound is inversely proportional to the minimum label set size

Avg Perceptron Mistake Bound under Linear Separability

Theorem

Let $(\mathbf{x}^1, Y^1), \ldots, (\mathbf{x}^T, Y^T)$ *be the examples presented to Avg Perceptron, where* $\mathbf{x}^t \in \mathbb{R}^d$ *and* $Y^t \subseteq [K]$. *Let* $W^* \in \mathbb{R}^{d \times K}$ $(\|W^*\| = 1)$ *be such that* $\frac{1}{|Y^t|} \sum_{i \in Y^t} \mathbf{w}_i^* . \mathbf{x}^t - \max_{j \in \overline{Y}^t} \mathbf{w}_j^* . \mathbf{x}^t \geq \gamma$, $\forall t \in [T]$. *Then we have:*

$$\sum_{t=1}^{T} L_A(h^t(\mathbf{x}^t), Y^t) \leq \frac{2}{\gamma^2} + \left[\frac{1}{c} + 1\right] \frac{R^2}{\gamma^2}$$

where $L_A(h(\mathbf{x}), Y) = \mathbb{I}_{\{h(\mathbf{x}) \notin Y\}}$, $c = \min_t |Y^t|$, $R = \max_t ||\mathbf{x}^t||$ *and* $\gamma \geq 0$ *is the margin of separation.*

- L_A: The 0-1 partial loss c: minimum label set size γ:margin
- The bound is inversely proportional to the minimum label set size

Avg Perceptron Mistake Bound under Non-Separability

Theorem

Let $(\mathbf{x}^1, Y^1), \ldots, (\mathbf{x}^T, Y^T)$ *be an input sequence presented to Avg Perceptron. Let* W $(\|W\| = 1)$ *be weight matrix corresponding to a multiclass classifier. Then for a fixed* $\gamma > 0$, *let* $d^t = \max\left\{0, \gamma - [\frac{1}{|Y^t|}\sum_{i \in Y^t} \mathbf{w}_i.\mathbf{x}^t - \max_{j \in \overline{Y}^t} \mathbf{w}_j.\mathbf{x}^t]\right\}$. *Let* $D^2 = \sum_{t=1}^{T}(|Y^t| d^t)^2$ *and* $R = \max_{t \in [T]} ||\mathbf{x}^t||$ *and* $c = \min_{t \in [T]} |Y^t|$. *Then, mistakes bound for Avg Perceptron is as follows.*

$$\sum_{t=1}^{T} L_A(h^t(\mathbf{x}^t), Y^t) \leq 2\frac{Z^2}{\gamma^2} + 2K\frac{R^2 + \Delta^2}{(\frac{\gamma}{Z})^2}$$

where $Z = \sqrt{1 + \frac{D^2}{\Delta^2}}$, $\Delta = \left[\frac{D^2 + KD^2R^2}{K}\right]^{\frac{1}{4}}$ *and* $K = \left[\frac{1}{c} + 1\right]$.

Avg Perceptron Mistake Bound under Non-Separability

Theorem

Let $(\mathbf{x}^1, Y^1), \ldots, (\mathbf{x}^T, Y^T)$ *be an input sequence presented to Avg Perceptron. Let* W $(\|W\| = 1)$ *be weight matrix corresponding to a multiclass classifier. Then for a fixed* $\gamma > 0$, *let* $d^t = \max\left\{0, \gamma - \left[\frac{1}{|Y^t|}\sum_{i \in Y^t} \mathbf{w}_i.\mathbf{x}^t - \max_{j \in \overline{Y}^t} \mathbf{w}_j.\mathbf{x}^t\right]\right\}$. *Let* $D^2 = \sum_{t=1}^{T}(|Y^t| d^t)^2$ *and* $R = \max_{t \in [T]} ||\mathbf{x}^t||$ *and* $c = \min_{t \in [T]} |Y^t|$. *Then, mistakes bound for Avg Perceptron is as follows.*

$$\sum_{t=1}^{T} L_A(h^t(\mathbf{x}^t), Y^t) \leq 2\frac{Z^2}{\gamma^2} + 2K\frac{R^2 + \Delta^2}{(\frac{\gamma}{Z})^2}$$

where $Z = \sqrt{1 + \frac{D^2}{\Delta^2}}$, $\Delta = \left[\frac{D^2 + KD^2R^2}{K}\right]^{\frac{1}{4}}$ *and* $K = \left[\frac{1}{c} + 1\right]$.

Avg Perceptron Mistake Bound under Non-Separability

Theorem

Let $(\mathbf{x}^1, Y^1), \ldots, (\mathbf{x}^T, Y^T)$ *be an input sequence presented to Avg Perceptron. Let* W $(\|W\| = 1)$ *be weight matrix corresponding to a multiclass classifier. Then for a fixed* $\gamma > 0$, *let* $d^t = \max\left\{0, \gamma - \left[\frac{1}{|Y^t|}\sum_{i \in Y^t} \mathbf{w}_i.\mathbf{x}^t - \max_{j \in \overline{Y}^t} \mathbf{w}_j.\mathbf{x}^t\right]\right\}$. *Let* $D^2 = \sum_{t=1}^{T}(|Y^t| d^t)^2$ *and* $R = \max_{t \in [T]} ||\mathbf{x}^t||$ *and* $c = \min_{t \in [T]} |Y^t|$. *Then, mistakes bound for Avg Perceptron is as follows.*

$$\sum_{t=1}^{T} L_A(h^t(\mathbf{x}^t), Y^t) \leq 2\frac{Z^2}{\gamma^2} + 2K\frac{R^2 + \Delta^2}{(\frac{\gamma}{Z})^2}$$

where $Z = \sqrt{1 + \frac{D^2}{\Delta^2}}$, $\Delta = \left[\frac{D^2 + KD^2R^2}{K}\right]^{\frac{1}{4}}$ *and* $K = \left[\frac{1}{c} + 1\right]$.

Avg Perceptron Mistake Bound under Non-Separability

Theorem

Let $(\mathbf{x}^1, Y^1), \ldots, (\mathbf{x}^T, Y^T)$ *be an input sequence presented to Avg Perceptron. Let* W $(\|W\| = 1)$ *be weight matrix corresponding to a multiclass classifier. Then for a fixed* $\gamma > 0$, *let* $d^t = \max\left\{0, \gamma - \left[\frac{1}{|Y^t|}\sum_{i \in Y^t} \mathbf{w}_i.\mathbf{x}^t - \max_{j \in \overline{Y}^t} \mathbf{w}_j.\mathbf{x}^t\right]\right\}$. *Let* $D^2 = \sum_{t=1}^{T}(|Y^t| d^t)^2$ *and* $R = \max_{t \in [T]} \|\mathbf{x}^t\|$ *and* $c = \min_{t \in [T]} |Y^t|$. *Then, mistakes bound for Avg Perceptron is as follows.*

$$\sum_{t=1}^{T} L_A(h^t(\mathbf{x}^t), Y^t) \leq 2\frac{Z^2}{\gamma^2} + 2K\frac{R^2 + \Delta^2}{(\frac{\gamma}{Z})^2}$$

where $Z = \sqrt{1 + \frac{D^2}{\Delta^2}}$, $\Delta = \left[\frac{D^2 + KD^2R^2}{K}\right]^{\frac{1}{4}}$ *and* $K = \left[\frac{1}{c} + 1\right]$.

Avg Pegasos and Max Pegasos

- **Objective Function** (at t): Sum of the Loss function and L2 regularizer of weights

Objective Function

$$f(W, \mathbf{x}^t, Y^t) = \frac{\lambda}{2}||W||^2 + L(h(\mathbf{x}^t), Y^t)$$

Avg Pegasos and Max Pegasos

- **Objective Function** (at t): Sum of the Loss function and L2 regularizer of weights
- λ: Regularization constant; $||W||$: Frobenius norm

Objective Function

$$f(W, \mathbf{x}^t, Y^t) = \frac{\lambda}{2}||W||^2 + L(h(\mathbf{x}^t), Y^t)$$

Avg Pegasos and Max Pegasos

- **Objective Function** (at t): Sum of the Loss function and L2 regularizer of weights
- λ: Regularization constant; $||W||$: Frobenius norm
- **Update** weights using gradient of objective function
  $\nabla^t = \nabla_{W^t} f(W^t, \mathbf{x}^t, Y^t)$

Objective Function

$$f(W, \mathbf{x}^t, Y^t) = \frac{\lambda}{2}||W||^2 + L(h(\mathbf{x}^t), Y^t)$$

Update

$$W^{t+1} = (W^t - \eta_t \nabla^t)$$

Project

$$W^{t+1} = \min\{1, \frac{1}{(\lambda||W^{t+1}||)}\}W^{t+1}$$

Avg Pegasos and Max Pegasos

- **Objective Function** (at t): Sum of the Loss function and L2 regularizer of weights
- λ: Regularization constant; $||W||$: Frobenius norm
- **Update** weights using gradient of objective function $\nabla^t = \nabla_{W^t} f(W^t, \mathbf{x}^t, Y^t)$
- $\nabla^t = \lambda W^t + \nabla_{W^t} L$ where the **sub-gradient** of the Avg Prediction or Max Prediction Loss is required

Objective Function

$$f(W, \mathbf{x}^t, Y^t) = \frac{\lambda}{2}||W||^2 + L(h(\mathbf{x}^t), Y^t)$$

Update

$$W^{t+1} = (W^t - \eta_t \nabla^t)$$

Project

$$W^{t+1} = \min\{1, \frac{1}{(\lambda||W^{t+1}||)}\}W^{t+1}$$

Avg Pegasos and Max Pegasos

- **Objective Function** (at t): Sum of the Loss function and L2 regularizer of weights
- λ: Regularization constant; $||W||$: Frobenius norm
- **Update** weights using gradient of objective function $\nabla^t = \nabla_{W^t} f(W^t, \mathbf{x}^t, Y^t)$
- $\nabla^t = \lambda W^t + \nabla_{W^t} L$ where the **sub-gradient** of the Avg Prediction or Max Prediction Loss is required
- Weights **projected** onto set B where $B = \{W : ||W|| \leq \frac{1}{\sqrt{\lambda}}\}$

Objective Function

$$f(W, \mathbf{x}^t, Y^t) = \frac{\lambda}{2}||W||^2 + L(h(\mathbf{x}^t), Y^t)$$

Update

$$W^{t+1} = (W^t - \eta_t \nabla^t)$$

Project

$$W^{t+1} = \min\{1, \frac{1}{(\lambda||W^{t+1}||)}\}W^{t+1}$$

Pegasos Regret Bound

- **Regret**: The *difference* between the loss incurred by your algorithm and the minimum loss incurred by a fixed policy (weight).

Pegasos Regret Bound

- **Regret**: The *difference* between the loss incurred by your algorithm and the minimum loss incurred by a fixed policy (weight).

Theorem

Let $(\mathbf{x}^1, Y^1), (\mathbf{x}^2, Y^1), \ldots, (\mathbf{x}^T, Y^T)$ *be an input sequence where* $\mathbf{x}^t \in \mathbb{R}^d$ *and* $Y^t \subseteq [K]$. *Let* $R = \max_t ||\mathbf{x}^t||$. *Then the* ***Regret*** *of Avg Pegasos is given as:*

$$\frac{1}{T}\sum_{t=1}^{T} f(W^t, \mathbf{x}^t, Y^t) - \min_W \frac{1}{T}\sum_{t=1}^{T} f(W, \mathbf{x}^t, Y^t) \leq \frac{G^2 ln T}{\lambda T}$$

where $G = \sqrt{\lambda} + \sqrt{1 + \frac{1}{c}}R$ *and* $c = \min_t |Y^t|$

Pegasos Regret Bound

- **Regret**: The *difference* between the loss incurred by your algorithm and the minimum loss incurred by a fixed policy (weight).

Theorem

Let $(\mathbf{x}^1, Y^1), (\mathbf{x}^2, Y^1), \ldots, (\mathbf{x}^T, Y^T)$ *be an input sequence where* $\mathbf{x}^t \in \mathbb{R}^d$ *and* $Y^t \subseteq [K]$. *Let* $R = \max_t ||\mathbf{x}^t||$. *Then the* ***Regret*** *of Avg Pegasos is given as:*

$$\frac{1}{T}\sum_{t=1}^{T} f(W^t, \mathbf{x}^t, Y^t) - \min_W \frac{1}{T}\sum_{t=1}^{T} f(W, \mathbf{x}^t, Y^t) \leq \frac{G^2 ln T}{\lambda T}$$

where $G = \sqrt{\lambda} + \sqrt{1 + \frac{1}{c}}R$ *and* $c = \min_t |Y^t|$

Experiments

- **Label Set Generation**: For all datasets, the candidate or partial label set for each instance contains the true label and some labels selected uniformly at random from the remaining labels
- We plot **Average Number of Misclassifications** (average 0-1 loss) with respect to the ***true label***
- **Baseline**: Results of Perceptron and Pegasos on true labels
- Results are averaged over 100 runs

Experiments

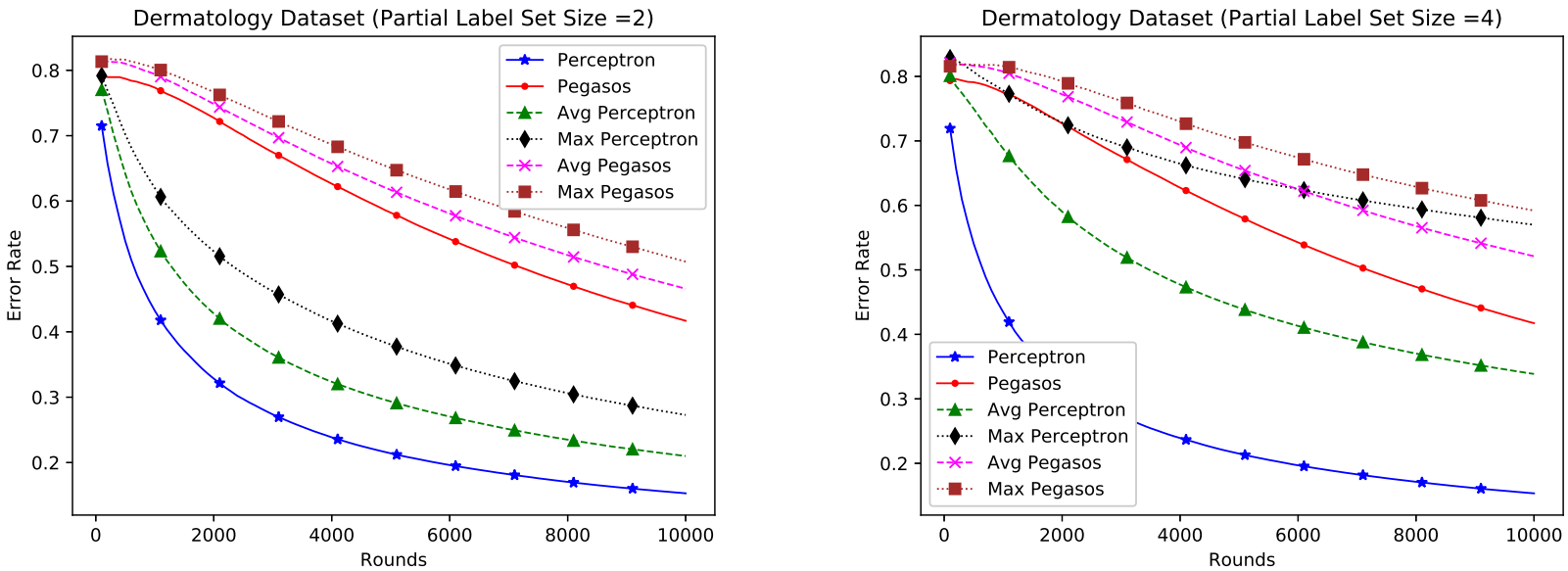


Figure 1: Dermatology Dataset (34 features, 6 classes)

Experiments

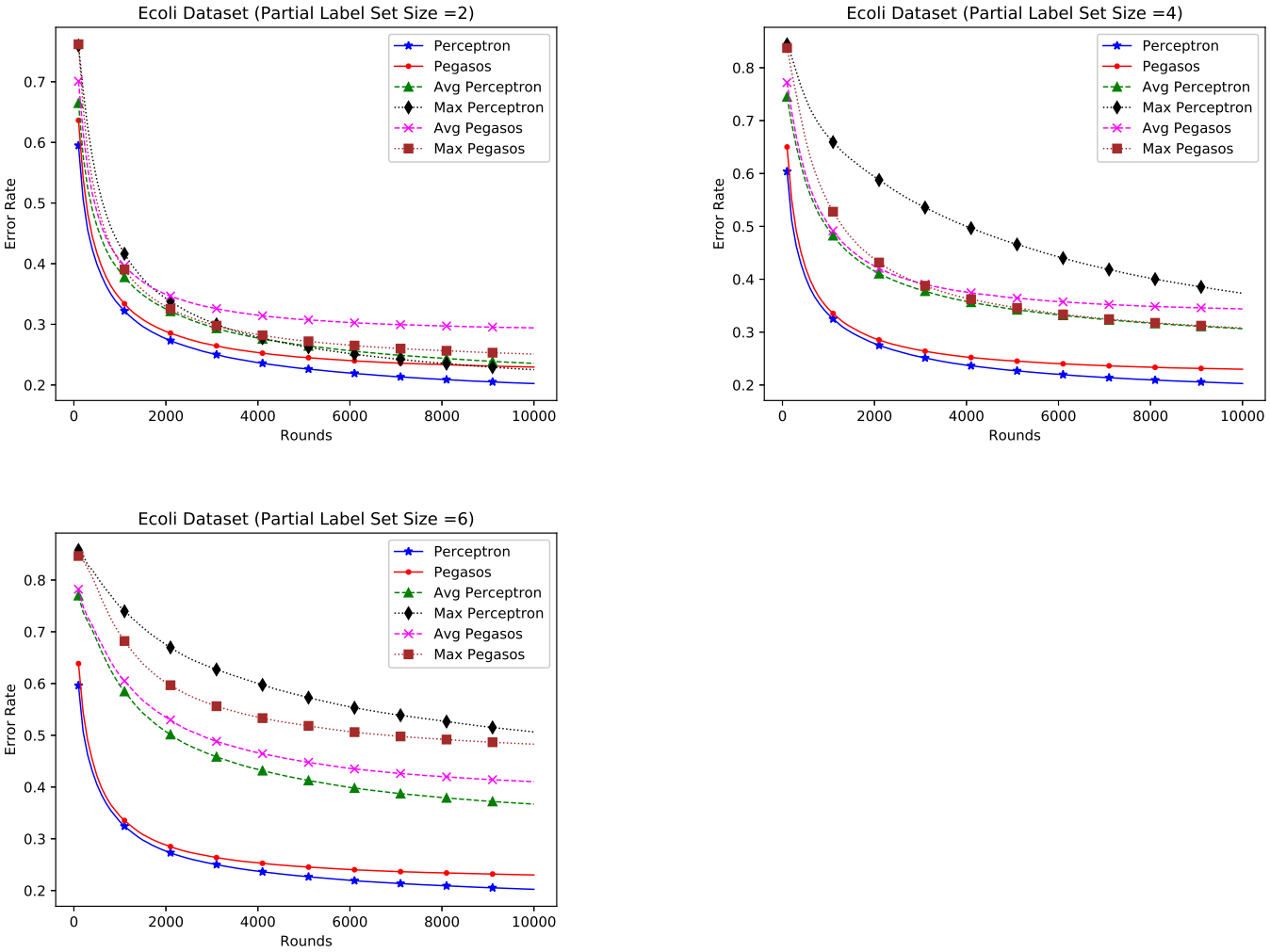


Figure 2: Ecoli Dataset (8 features, 8 classes)

Experiments

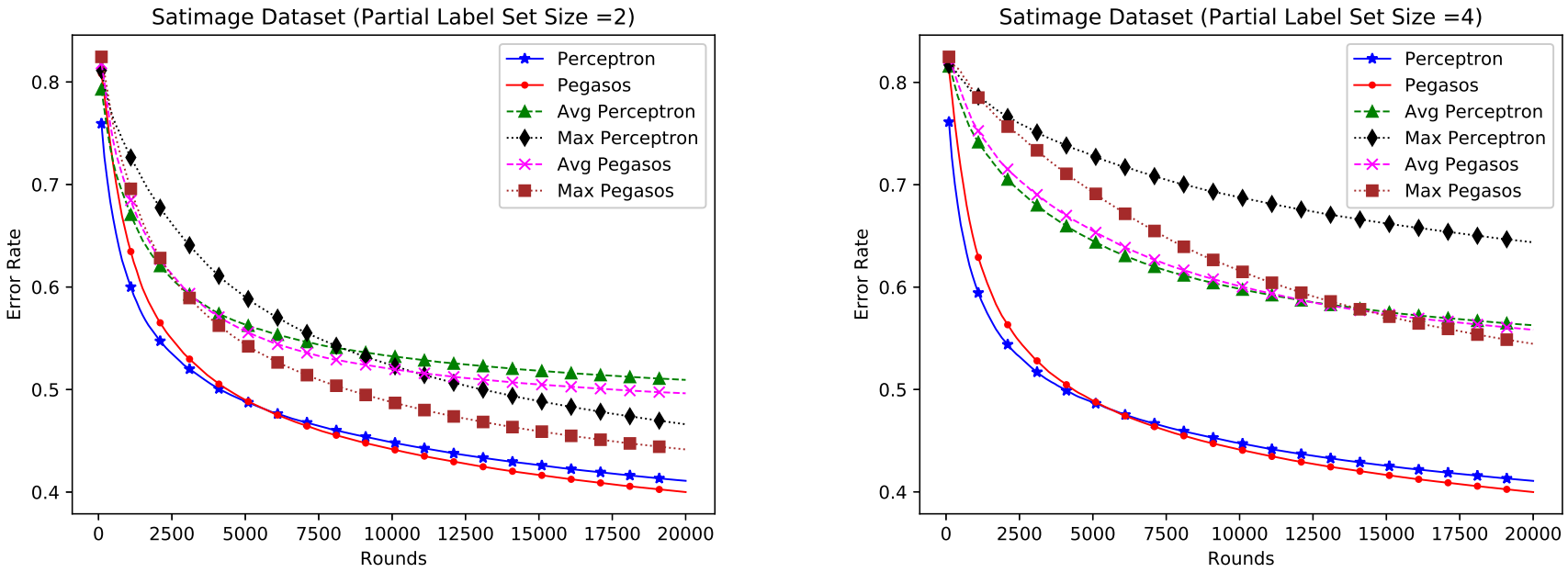


Figure 3: Satimage Dataset (36 features, 6 classes)

Experiments

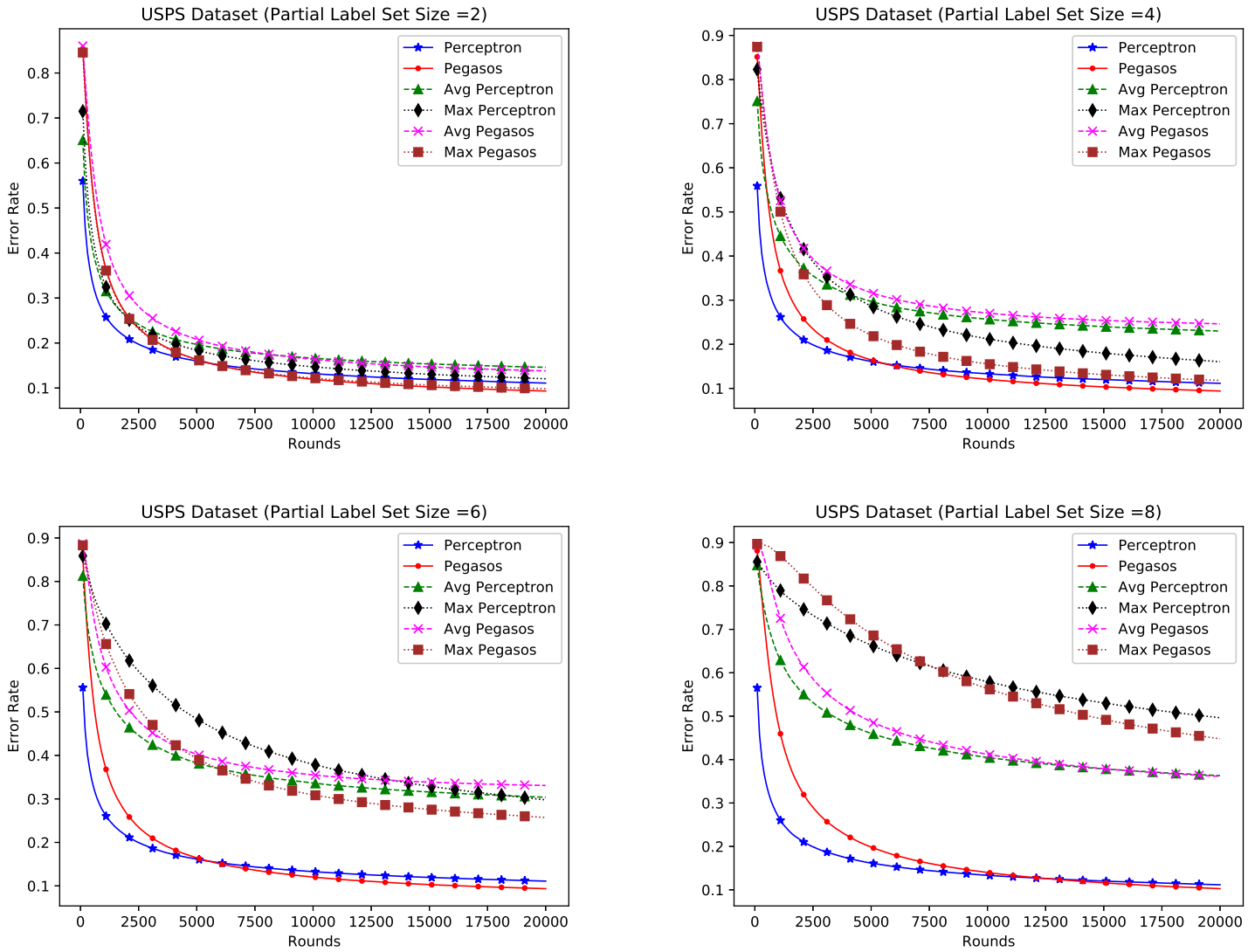


Figure 4: USPS Dataset (256 features, 10 classes)

Experiments

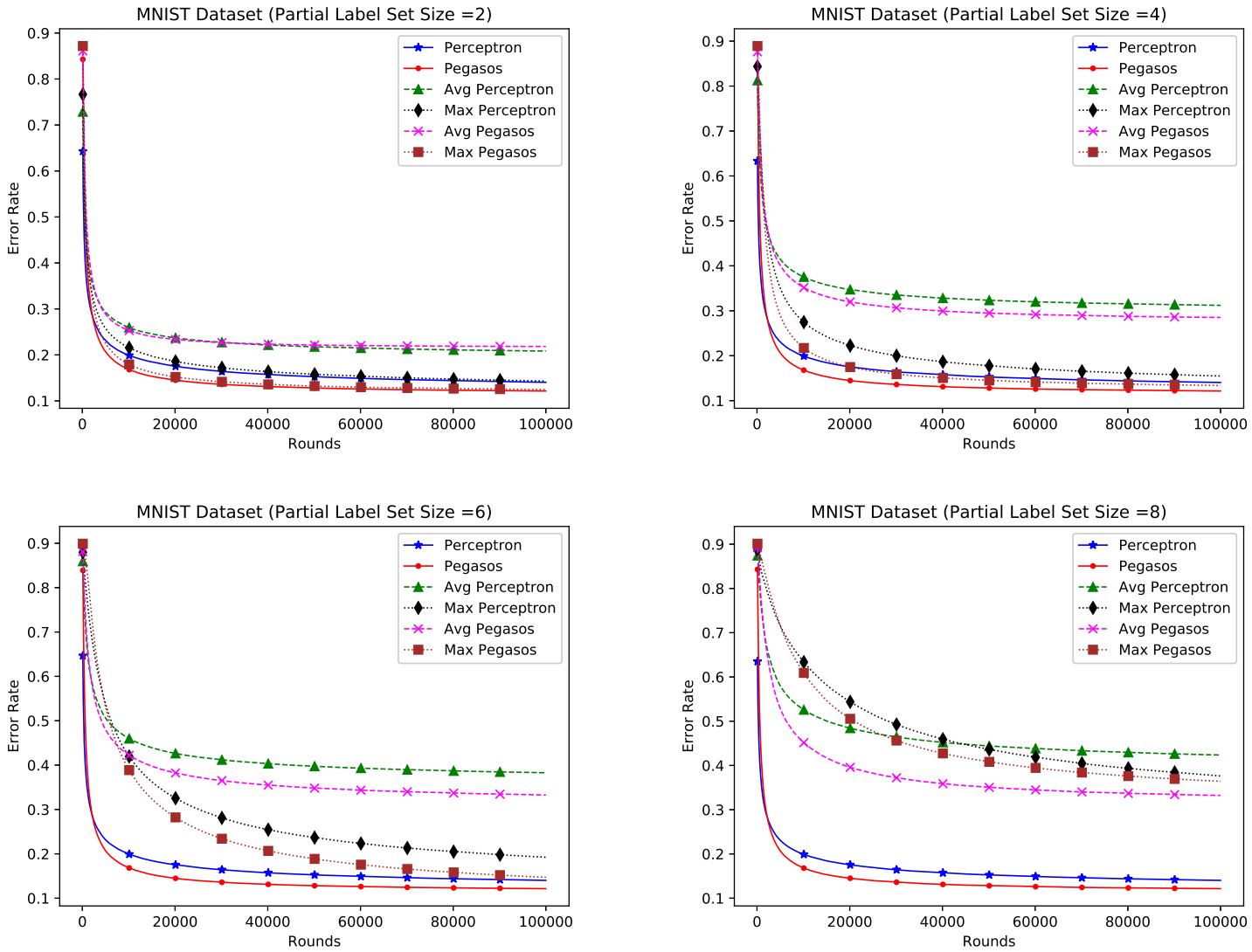


Figure 5: MNIST Dataset (784 features, 10 classes)

Summary

- We proposed Perceptron and Pegasos based algorithms for multiclass classification with Partial Labels

Summary

- We proposed Perceptron and Pegasos based algorithms for multiclass classification with Partial Labels
- Loss functions used are Average Prediction (convex) and Max Prediction Loss (non-convex)

Summary

- We proposed Perceptron and Pegasos based algorithms for multiclass classification with Partial Labels
- Loss functions used are Average Prediction (convex) and Max Prediction Loss (non-convex)
- Mistake and Regret bounds for Average Prediction Loss

Summary

- We proposed Perceptron and Pegasos based algorithms for multiclass classification with Partial Labels
- Loss functions used are Average Prediction (convex) and Max Prediction Loss (non-convex)
- Mistake and Regret bounds for Average Prediction Loss
- Max Prediction Loss gives better results for small label set sizes

Summary

- We proposed Perceptron and Pegasos based algorithms for multiclass classification with Partial Labels
- Loss functions used are Average Prediction (convex) and Max Prediction Loss (non-convex)
- Mistake and Regret bounds for Average Prediction Loss
- Max Prediction Loss gives better results for small label set sizes
- *Next*: A theoretical analysis to understand the performance Max Prediction Loss

Thank You