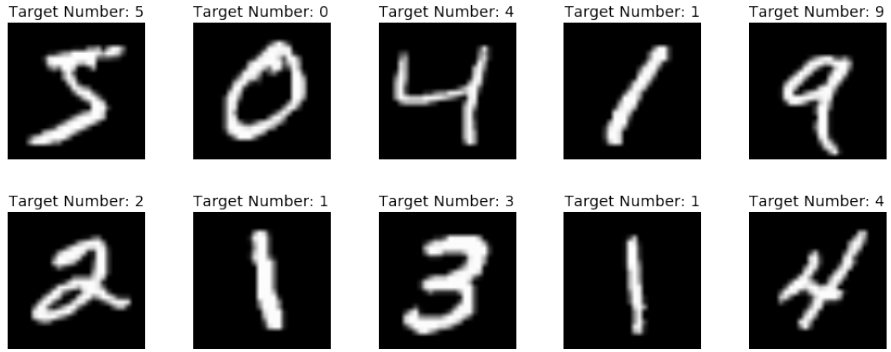
COS324: Introduction to Machine Learning

Lecture 11: Multiclass Problems

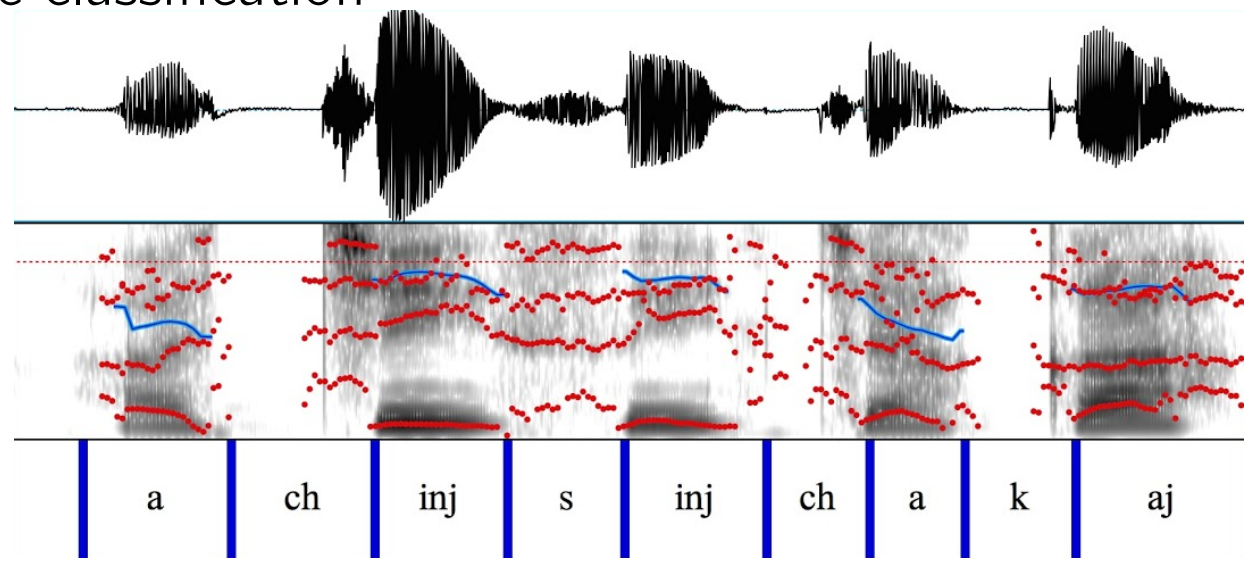
Prof. Elad Hazan & Prof. Yoram Singer

Examples for Multiclass Problems

- Digit recognition



- Phoneme classification



Multiclass: Problem Setting

- Instances: $\mathbf{x} \in \mathbb{R}^d$
- Labels: $y \in [k] = \{1, \ldots, k\}$
- Multiclass predictor: $h : \mathbb{R}^d \to [k]$
- Classification error:

$$\mathbb{1}[h(\mathbf{x}) \neq y]$$

- As in binary case, minimizing multiclass prediction error is typically computationally NP-hard

One-vs-Rest / One-vs-All

- Learn k binary predictors $h_j : \mathbb{R}^d \to \mathbb{R}$
  (not $h_j : \mathbb{R}^d \to \{-1, +1\}$)
- j'th predictor distinguishes j'th class from the rest
- Learning scheme
  I. Transform $S \mapsto S^1, S^2, \ldots, S^k$ where

$$S^j = \left\{ \left(\mathbf{x}_i, (-1)^{\mathbb{1}[y_i \neq j]} \right) \right\}_{i=1}^{m}$$

  II. For $j = 1, \ldots, k$ call $\mathcal{A}(S^j) \mapsto h_j(\cdot)$
- Inference:

$$\hat{y} = \arg\max_{j \in [k]} h_j(\mathbf{x})$$

Example

- Label transformation table for a 4 class problem

y	y^1	y^2	y^3	y^4
1	+	-	-	-
2	-	+	-	-
3	-	-	+	-
4	-	-	-	+

- Original training set $S = \{(\mathbf{x}_1, 2), (\mathbf{x}_2, 4), (\mathbf{x}_3, 2), (\mathbf{x}_4, 3), (\mathbf{x}_5, 1)\}$

- Results in the following 4 binary problems

S^1	S^2	S^3	S^4
$(\mathbf{x}_1, -)$	$(\mathbf{x}_1, +)$	$(\mathbf{x}_1, -)$	$(\mathbf{x}_1, -)$
$(\mathbf{x}_2, -)$	$(\mathbf{x}_2, -)$	$(\mathbf{x}_2, -)$	$(\mathbf{x}_2, +)$
$(\mathbf{x}_3, -)$	$(\mathbf{x}_3, +)$	$(\mathbf{x}_3, -)$	$(\mathbf{x}_3, -)$
$(\mathbf{x}_4, -)$	$(\mathbf{x}_4, -)$	$(\mathbf{x}_4, +)$	$(\mathbf{x}_4, -)$
$(\mathbf{x}_5, +)$	$(\mathbf{x}_5, -)$	$(\mathbf{x}_5, -)$	$(\mathbf{x}_5, -)$

Linear Multiclass Predictors

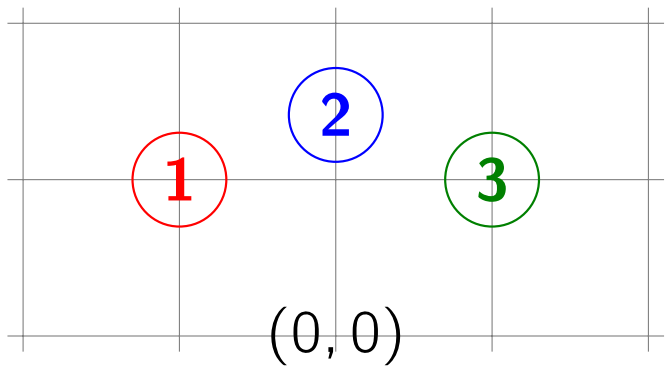
- Assume $h_j(\mathbf{x}) = \mathbf{w}_j \cdot \mathbf{x}$ where $\mathcal{A}(S^j) \mapsto h_j$
- Construct matrix W of size $k \times d$ whose j'th row is $\mathbf{w}_j$

$$W = \begin{bmatrix} -\mathbf{w}_1 - \\ -\mathbf{w}_2 - \\ \cdots \\ -\mathbf{w}_k - \end{bmatrix}$$

- Predicted scores $\mathbf{z} = W\mathbf{x}$
- Predicted class $\hat{y} = \arg\max_{j \in [k]} z_j$

Deficiency of One-vs-All

- Predictors are trained independently
- Resulting binary problems may be overly difficult
- One-vs-All would fail in the following setting



- Problem is linearly separable

$$W = \begin{bmatrix} -1 & 1 \\ 0 & \sqrt{2} \\ 1 & 1 \end{bmatrix}$$

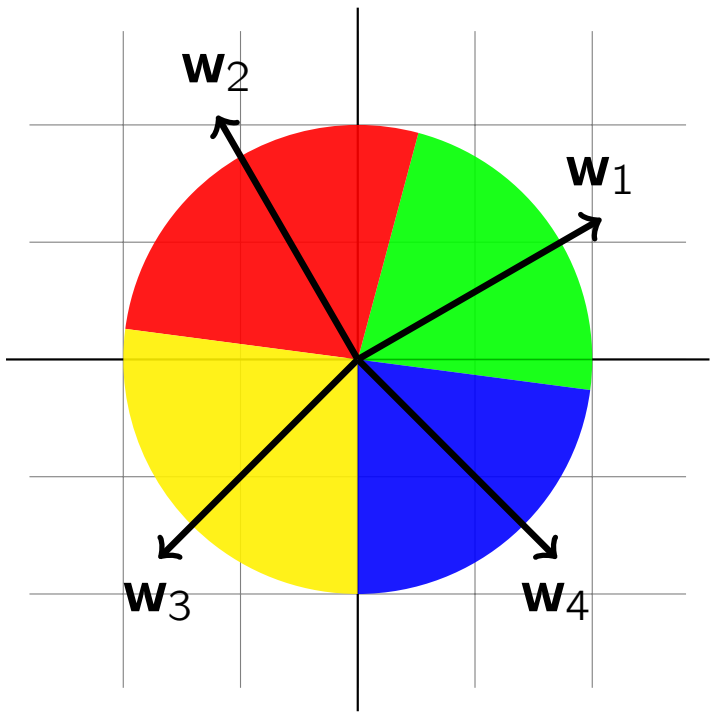
Multiclass Margin

- Impose directly

$$(\mathbf{x}, y) \quad \Rightarrow \quad \mathbf{w}_y \cdot \mathbf{x} > \mathbf{w}_j \cdot \mathbf{x} \quad (j \neq y)$$

or in matrix-vector format

$$(\mathbf{x}, y) \quad \Rightarrow \quad [W\mathbf{x}]_y > [W\mathbf{x}]_j \quad (j \neq y)$$



Margin Loss

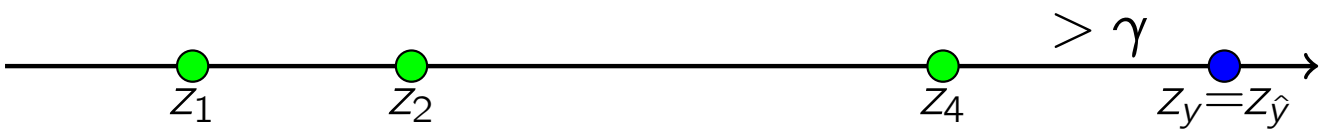
- Abbreviate
$$z_j \stackrel{\text{def}}{=} \mathbf{w}_j \cdot \mathbf{x}$$
- Predicted label
$$\hat{y} = \arg\max_{j \in [k]} z_j$$
- Multiclass prediction error error
$$\ell^{\text{MC}}(\mathbf{z}) \stackrel{\text{def}}{=} \mathbb{1}[\hat{y} \neq y]$$
- Finding W which minimizes multiclass error is computationally hard and thus we use convex surrogate loss

Max-Margin Loss

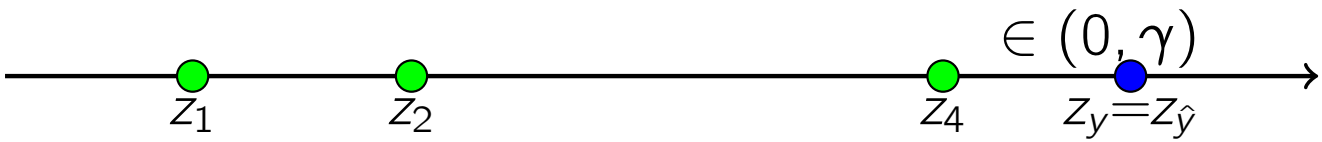
- Difference of scores + additional penalty γ

$$\ell^{\text{MM}}(\mathbf{z}) \stackrel{\text{def}}{=} \left[\gamma + \max_{j \neq y} z_j - z_y\right]_+$$

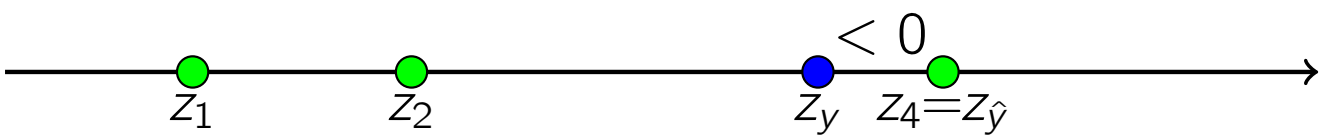
- Margin great than $\gamma \;\Rightarrow\; \ell^{\text{MC}} = \ell^{\text{MM}} = 0$



- Margin $\in (0, \gamma) \;\Rightarrow\; \ell^{\text{MC}} = 0$ but $\ell^{\text{MM}} \geq 0$

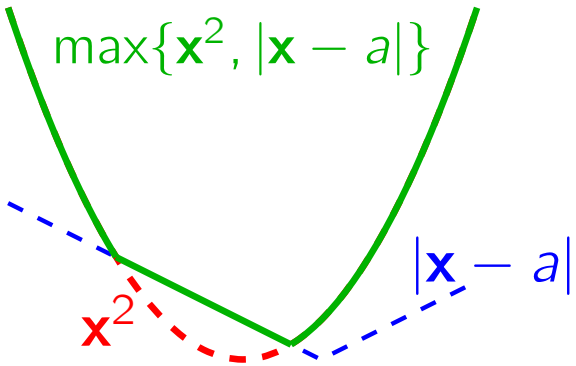


- Margin $< 0 \;\Rightarrow\; \ell^{\text{MC}} = 1$ and $\ell^{\text{MM}} \geq \gamma$



Convexity of Max-Margin Loss

- $\mathbf{w}_j \cdot \mathbf{x}$ is linear in $\mathbf{w}_j$ hence both convex and concave
- Maximum of convex functions is convex (prove)



hence $\psi(W) \stackrel{\text{def}}{=} \max_{j \neq y} \mathbf{w}_j \cdot \mathbf{x}$ is convex in W

- Sum of convex functions is convex hence ϕ convex

$$\phi(W) \stackrel{\text{def}}{=} \gamma + \psi(W) - \mathbf{w}_y \cdot \mathbf{x}$$

- Max-margin is maximum of two convex functions hence convex

$$\ell^{\text{MM}}(W) \stackrel{\text{def}}{=} [\phi(W)]_+ \stackrel{\text{def}}{=} \max\{0, \phi(W)\}$$

Multivariate Logistic Regression

- As before $z_j = \mathbf{w}_j \cdot \mathbf{x}$ or in matrix notation $z_j = [W\mathbf{x}]_j$
- Define the "probability" label is i as

$$\mathbb{P}[i|\mathbf{x}] = \frac{\exp(z_i)}{\sum_{j=1}^{k} \exp(z_j)}$$

- Loss is negative log-probability of correct class

$$\begin{aligned}\ell^{\mathrm{LR}}(\mathbf{z}) &= -\log(\mathbb{P}[y|\mathbf{x}]) \\ &= \underbrace{\log\left(\textstyle\sum_j \exp(z_j)\right)}_{\mathrm{SoftMax}} - z_y\end{aligned}$$

- Denote $\hat{y} = \arg\max_j z_j$ and if $\hat{y} \neq y$ then

$$\ell^{\mathrm{LR}}(\mathbf{z}) \geq \log \underbrace{\left(\exp(z_{\hat{y}}) + \exp(z_y)\right)}_{\geq 2\exp(z_y)} - z_y \geq \log(2)$$

- Therefore $\ell^{\mathrm{MC}} = 1 \;\Rightarrow\; \ell^{\mathrm{LR}} \geq \log(2)$

Yet Another Reminder of SGD

- Parameters: ρ, η^0, T_0, T
- Initialize $\mathbf{w}^1 = \mathbf{0}$
- For $t = 1, \ldots, T$:
  - Set learning-rate $\eta^t = \frac{\eta^0}{t}$
  - Choose $S' \subset S$ at random
  - Update

$$\mathbf{w}^{t+1/2} = \mathbf{w}^t - \eta^t \nabla \hat{f}(\mathbf{w}^t) \; = \mathbf{w}^t - \eta^t \frac{1}{|S'|} \sum_{i \in S'} \nabla f_i(\mathbf{w}^t)$$

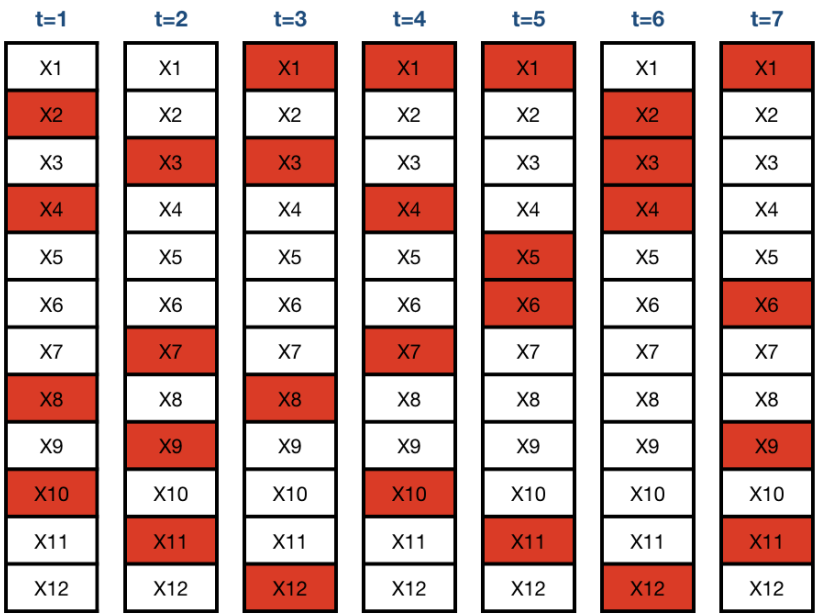
  - Project onto $\Omega = \{\mathbf{w} \mid \|\mathbf{w}\| \leq \rho\}$

$$\mathbf{w}^{t+1} = \min \left\{ 1, \frac{b}{\|\mathbf{w}^{t+1/2}\|} \right\} \mathbf{w}^{t+1/2}$$

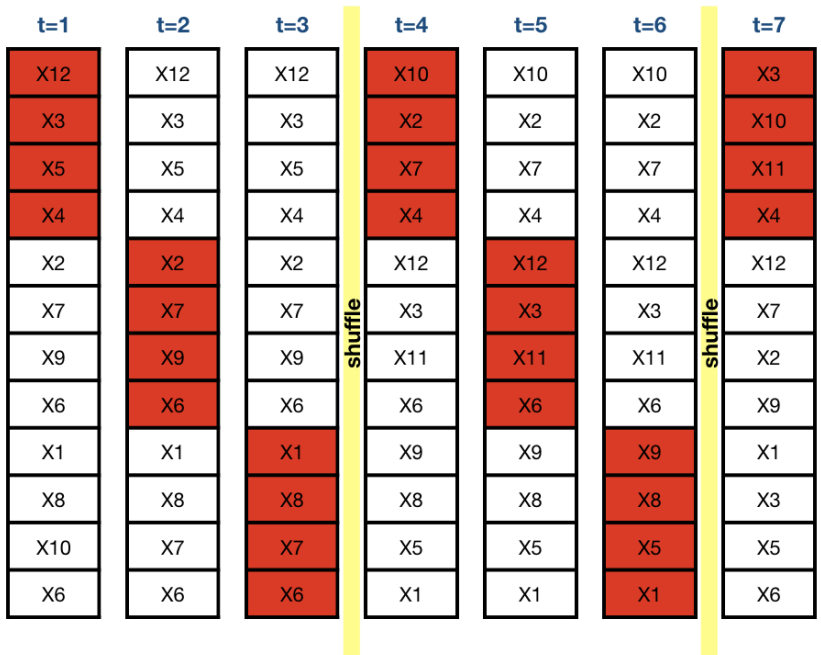
- Output $\bar{\mathbf{w}}^T = \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \mathbf{w}^t$ // Suffix-averaging

Mini-batching

With Replacement



No Replacement



SGD for Multiclass

- Instead of vector $\mathbf{w}^t$ keep matrix W^t
- Since loss is a function of W^t the gradient a matrix

$$\nabla \hat{f}(W^t) \in \mathbb{R}^{k \times d}$$

- Gradient step (before projection)

$$W^{t+1/2} = W^t - \eta^t \nabla \hat{f}(W^t)$$

- Projection accounts for matrix structure; W^{t+1} satisfies

$$\begin{bmatrix} \|-\mathbf{w}_1-\| \leq \rho \\ \|-\mathbf{w}_2-\| \leq \rho \\ \cdots \\ \|-\mathbf{w}_k-\| \leq \rho \end{bmatrix}$$

- Prove: all matrices with norm of each row is bounded by ρ define a convex set

Multiclass Logistic Regression

- Calculate $z = W^t \mathbf{x}^t$
- Calculate

$$p_j^t = \frac{\exp(z_j)}{\sum_{i=1}^k \exp(z_i)} \quad \text{and then} \quad p_y^t \mathrel{-}= 1$$

- Update

$$W^{t+1/2} = W^t - \eta^t \mathbf{p}^t \left(\mathbf{x}^t\right)^\top$$

  which amounts to

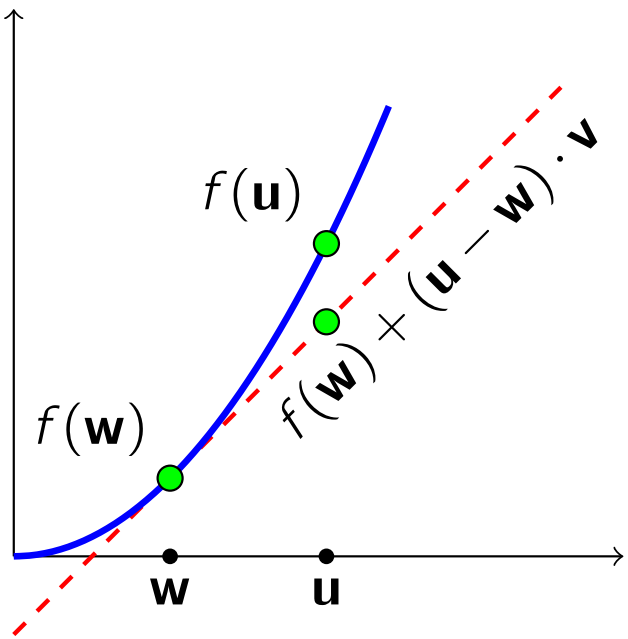
$$\mathbf{w}_j^{t+1/2} = \mathbf{w}_j^t - \eta^t p_j^t \mathbf{x}^t$$

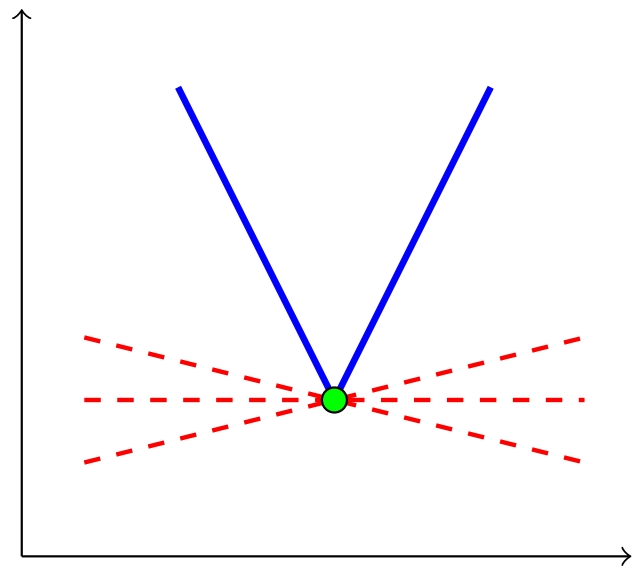
- Project

$$\mathbf{w}_j^{t+1} = \min\left\{1, \frac{\rho}{\|\mathbf{w}_j^{t+1/2}\|}\right\} \mathbf{w}_j^{t+1/2}$$

Sub-gradients

- $\mathbf{v}$ is sub-gradient of f at $\mathbf{w}$ if $\forall \mathbf{u}, \quad f(\mathbf{u}) \geq f(\mathbf{w}) + \mathbf{v} \cdot (\mathbf{u} - \mathbf{w})$
- Differential set, $\partial f(\mathbf{w})$, is the set of sub-gradients of f at $\mathbf{w}$





SGD for Max-Margin Multiclass

- Find labels with margin error

$$\mathcal{E} = \left\{ j \neq y \,|\, \gamma + z_j - z_y \geq 0 \right\}$$

- If $\mathcal{E} = \emptyset$ set $W^{t+1} = W^t$

- Choose $\mathbf{p}^t_{\downarrow y} \in \Delta_{|\mathcal{E}|}$ and $p^t_y = 1$

- Update

$$W^{t+1/2} = W^t - \eta^t \, \mathbf{p}^t \left(\mathbf{x}^t\right)^\top \quad \Leftrightarrow \quad W^{t+1/2} = W^t - \eta^t \begin{bmatrix} p^t_1 \, \mathbf{w}_1 \\ p^t_2 \, \mathbf{w}_2 \\ \cdots \\ p^t_k \, \mathbf{w}_k \end{bmatrix}$$

- MM-update chooses a sub-gradient from differential set

$$\partial_W \left[\gamma + \max_{j \neq y} [Wx]_j - [Wx]_y \right]_+$$

Examples and Elaboration

- Let $\tilde{y} = \arg\max_{j \in \mathcal{E}} z_j$
- Update of y'th row of $W^{t+1/2}$

$$\mathbf{w}_y^{t+1/2} = \mathbf{w}_y^t + \eta^t \mathbf{x}^t$$

- Max-violator

$$\mathbf{w}_{\tilde{y}}^{t+1/2} = \mathbf{w}_{\tilde{y}}^t - \eta^t \mathbf{x}^t$$

- Uniform weights of violators

$$\forall j \in \mathcal{E} : \ \mathbf{w}_j^{t+1/2} = \mathbf{w}_j^t - \frac{\eta^t}{|\mathcal{E}|} \mathbf{x}^t$$

- Margin-based weighted violators

$$\forall j \in \mathcal{E} : \ p_j = \frac{\gamma + z_j - z_y}{\sum_{i \in \mathcal{E}} \gamma + z_i - z_y} \qquad \mathbf{w}_j^{t+1/2} = \mathbf{w}_j^t - \eta^t p_j \, \mathbf{x}^t$$

Max-margin vs. Logistic Regression

- Both updates of the form

$$\mathbf{w}_j^{t+1/2} = \mathbf{w}_j^t - \eta^t p_j \mathbf{x}^t$$

- Both updates satisfy

$$\sum_j p_j = 0$$

- If $\mathcal{E} \neq \emptyset$ for MM $p_y = -1$ and for LR $p_y > -1$
- If $\hat{y}^t = y^t$ and $\mathcal{E} = \emptyset$ $\mathbf{p}^t = \mathbf{0}$ then for MM and $\mathbf{p}^t \approx \mathbf{0}$ for LR
- LR is a dense update $|\{j : p_j^t > 0\}| = k - 1$
- MM is sparse update $|\{j : p_j^t > 0\}| \leq |\mathcal{E}|$

Cost-Sensitive Multiclass

- Classes often have semantic meaning and similarities
- Image classification: **Ape** $\approx$ **Baboon** but **Ape** $\not\approx$ **Subaru**


- Associate cost with confusing class y with $\hat{y}$,

$$\forall y \neq y' \ \Delta(y, y') > 0 \quad \text{and} \quad \forall y \ \Delta(y, y) = 0$$

- Replace a fixed margin of γ with label-dependent margin

Cost-Sensitive Multiclass

- Prediction mechanism stays intact

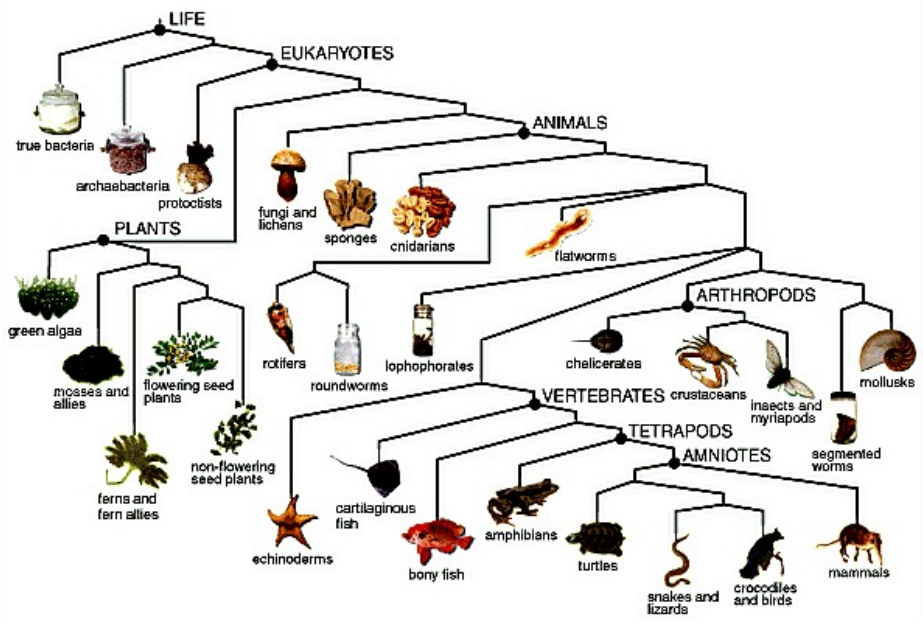
$$\hat{y} = \max_{j=1}^{k} \underbrace{[W\mathbf{x}]_j}_{\stackrel{\text{def}}{=} z_j}$$

- Convex surrogate for cost $\Delta(y, \hat{y})$

$$\begin{aligned} \Delta(y, \hat{y}) &\leq \Delta(y, \hat{y}) + z_{\hat{y}} - z_y \\ &\leq \underbrace{\max_r \Delta(r, \hat{y}) + z_r - z_y}_{\ell(y, \mathbf{z})} \end{aligned}$$

Use Exampe: Hierarchical Categories

- Classes are organized in a tree



- Length of (unique) path from y to $\hat{y}$ is $\Delta(y, \hat{y})$

$$\Delta(\text{turtles,snakes}) = 1 \quad \Delta(\text{bacteria,mammals}) = 14 \quad \ldots$$

- Replace γ with path length $\Delta(\cdot, \cdot)$

- Use max-violation update

Summary

- Multiclass categorization setting
- One-vs-rest decomposition
- Margin losses: max-margin & log-loss
- Gradient-based learning of multiclass predictors
- Cost-sensative max-margin