Online Knowledge-Based Support Vector Machines

Gautam Kunapuli[1], Kristin P. Bennett[2], Amina Shabbeer[2], Richard Maclin[3] and Jude W. Shavlik[1]

[1]University of Wisconsin-Madison, USA
[2]Rensselaer Polytechnic Institute, USA
[3]University of Minnesota, Duluth, USA

ECML 2010, Barcelona, Spain

Outline

- ***Knowledge-Based Support Vector Machines***
- The Adviceptron: Online KBSVMs
- A Real-World Task: Diabetes Diagnosis
- A Real-World Task: Tuberculosis Isolate Classification
- Conclusions

Knowledge-Based SVMs

- Introduced by Fung et al (2003)
- Allows incorporation of expert advice into SVM formulations
- Advice is specified with respect to polyhedral regions in input (feature) space

$$(\texttt{feature}_7 \geq 5) \wedge (\texttt{feature}_{12} \leq 4) \Rightarrow (\texttt{class} = +1)$$
$$(\texttt{feature}_2 \leq -3) \wedge (\texttt{feature}_3 \leq 4) \wedge (\texttt{feature}_{10} \geq 0) \Rightarrow (\texttt{class} = -1)$$
$$(3\texttt{feature}_6 + 5\texttt{feature}_8 \geq 2) \wedge (\texttt{feature}_{11} \leq -3) \Rightarrow (\texttt{class} = +1)$$

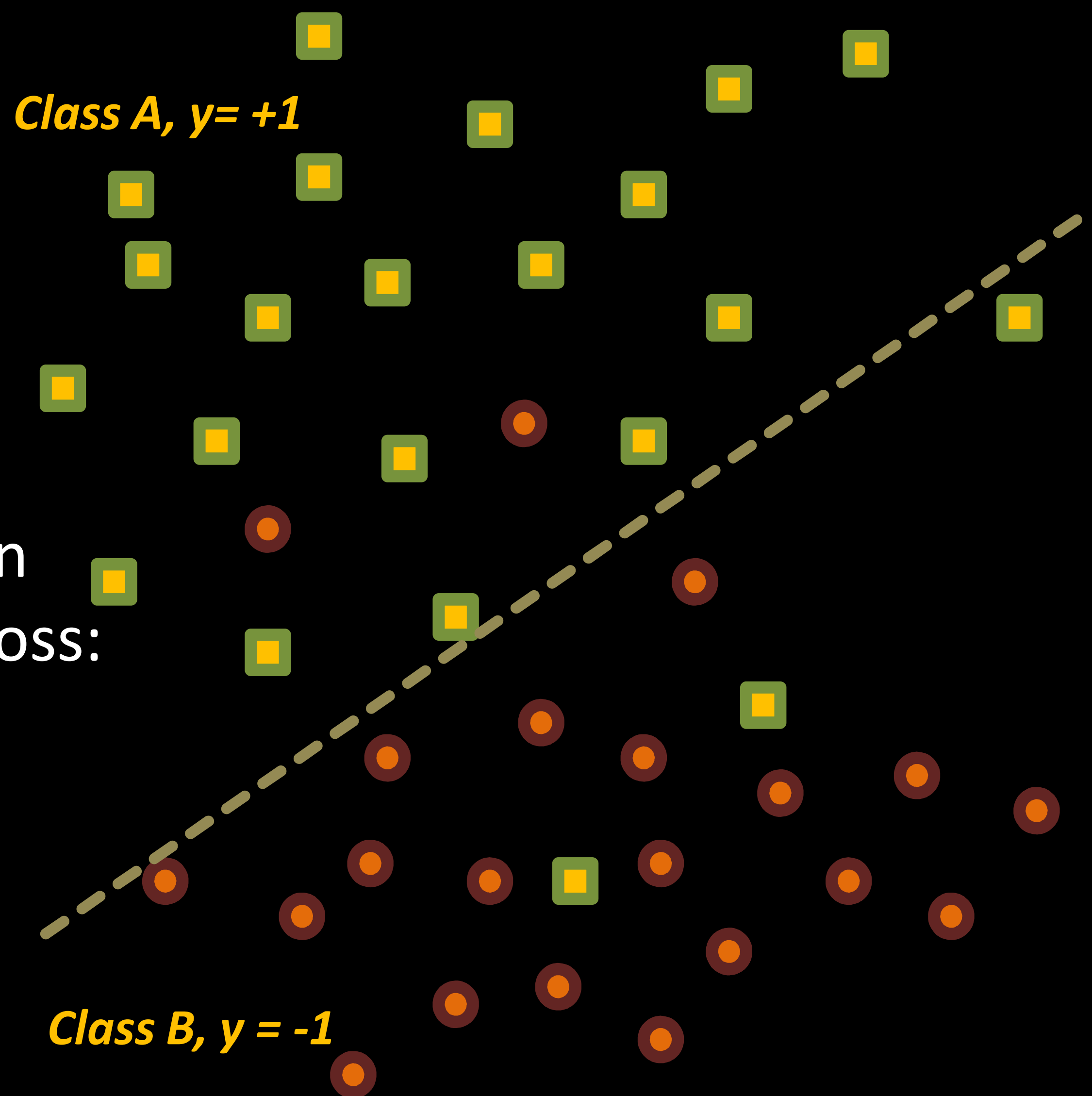
- Can be incorporated into SVM formulation as constraints using ***advice variables***

Knowledge-Based SVMs

In classic SVMs, we have *T* labeled data points $(\mathbf{x}^t, y_t)$, $t = 1, \ldots, T$. We learn a linear classifier $\mathbf{w}'\mathbf{x} - b = 0$.

The standard SVM formulation trades off regularization and loss:

$$\begin{aligned} \min \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \lambda\,\mathbf{e}'\xi \\ \text{sub. to} \quad & Y(X\mathbf{w} - b\mathbf{e}) + \xi \geq \mathbf{e}, \\ & \xi \geq 0. \end{aligned}$$



Knowledge-Based SVMs

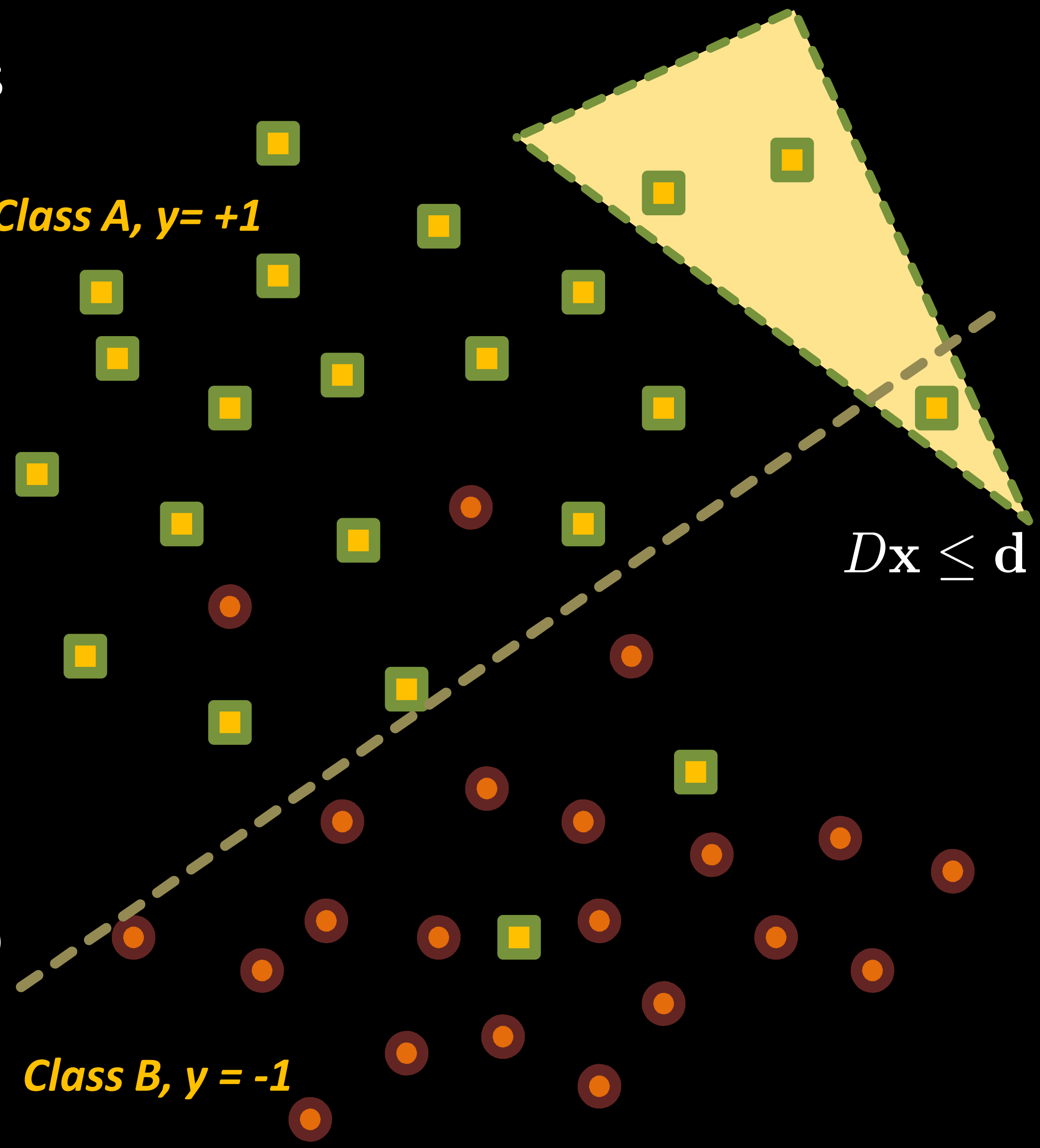
We assume an expert provides polyhedral advice of the form

$$D\mathbf{x} \leq \mathbf{d} \Rightarrow \mathbf{w}'\mathbf{x} \geq b$$

We can transform the logic constraint above using ***advice variables, u***

$$\begin{aligned} D'\mathbf{u} + \mathbf{w} &= 0, \\ -\mathbf{d}'\mathbf{u} - b &\geq 0, \\ \mathbf{u} &\geq 0 \end{aligned}$$

These constraints are added to the standard formulation to give ***Knowledge-Based SVMs***



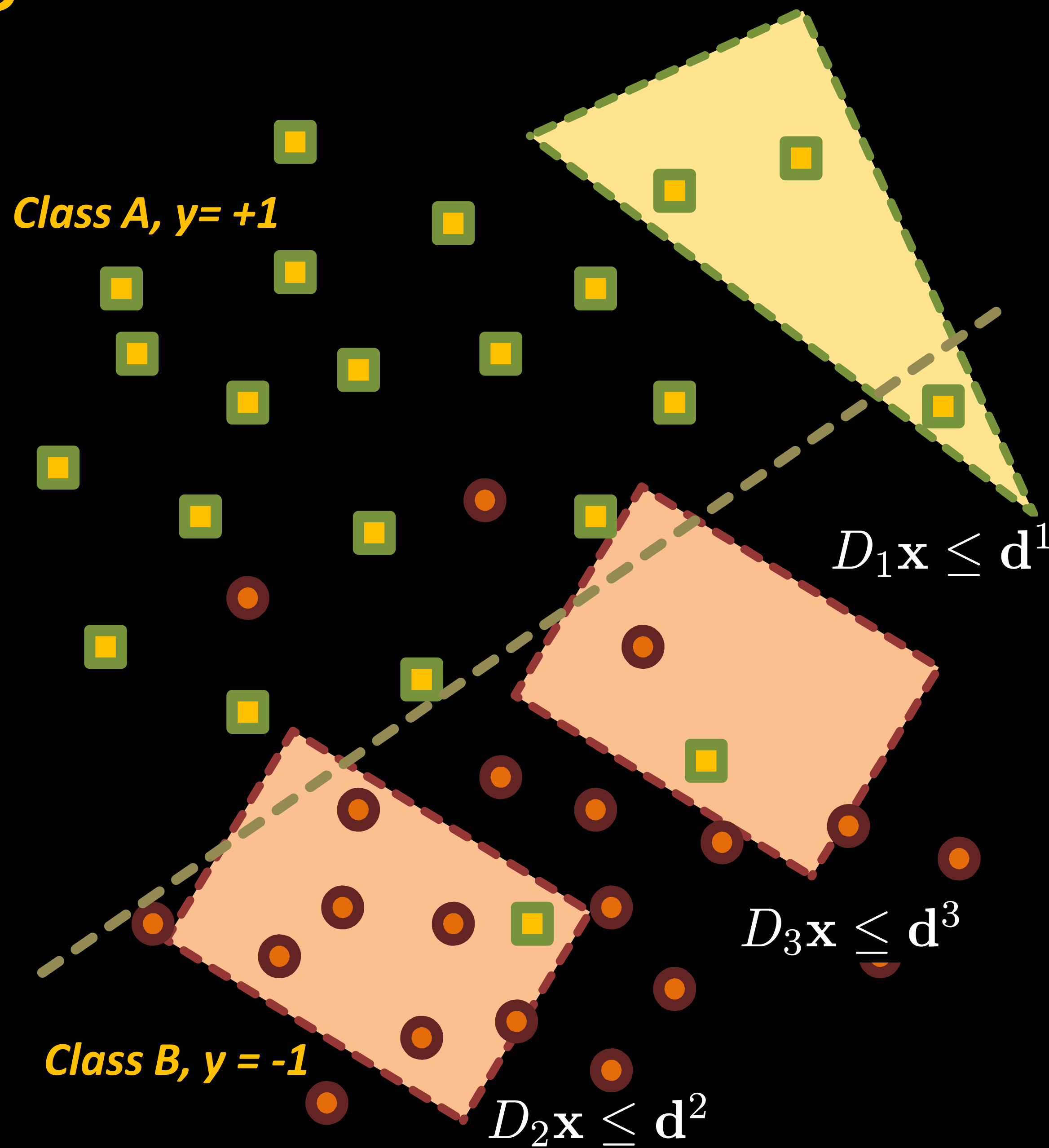
Knowledge-Based SVMs

In general, there are ***m* advice sets**, each with label $z = \pm 1$ for advice belonging to Class A or B,

$$D_i \mathbf{x} \leq \mathbf{d}^i \Rightarrow z_i\left(\mathbf{w}'\mathbf{x}\right) - b \geq 0$$

Each advice set ***adds the following constraints*** to the SVM formulation

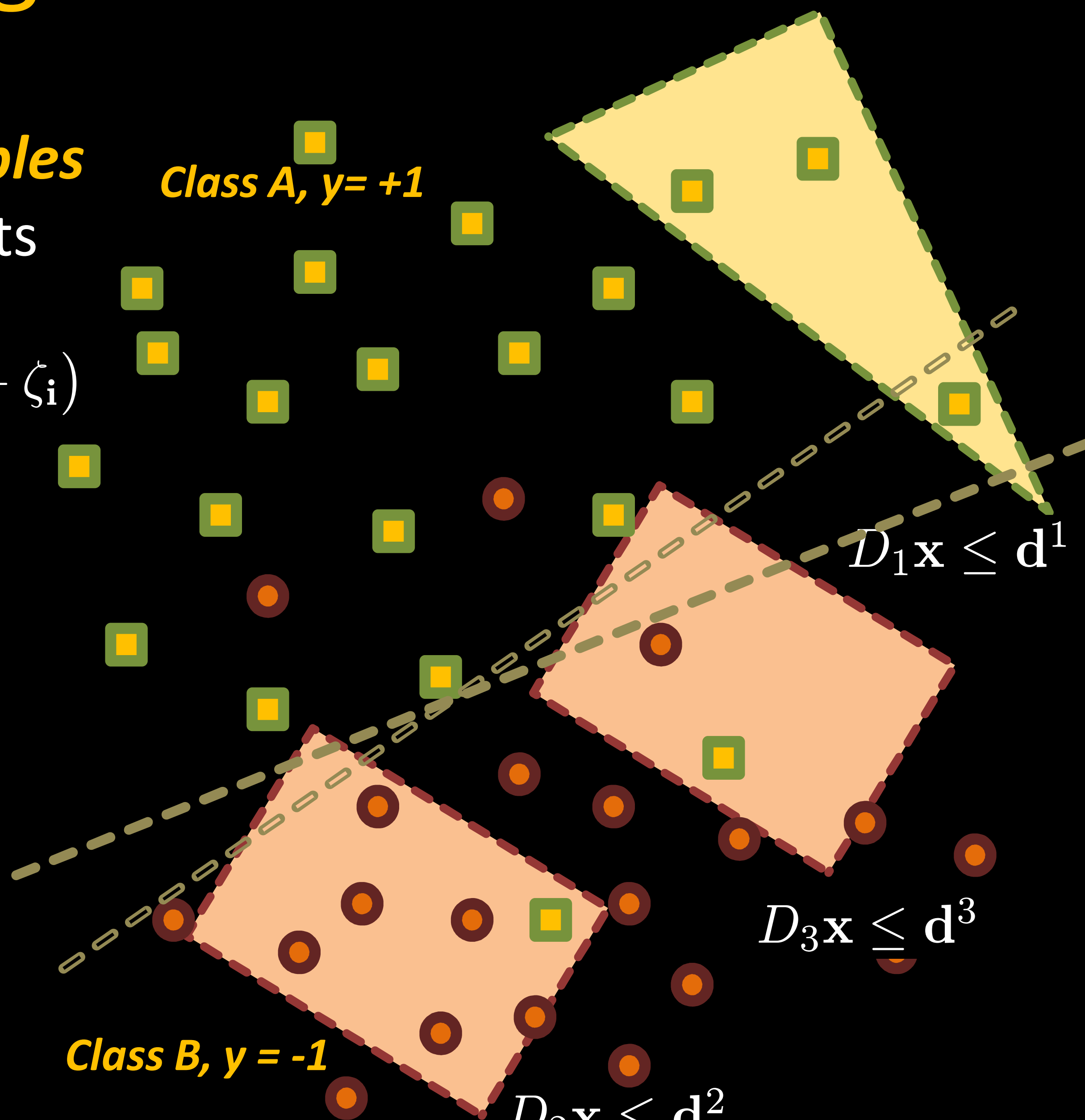
$$\begin{aligned} D_i' \mathbf{u}^i + z_i \mathbf{w} &= 0, \\ -\mathbf{d}^{i\prime} \mathbf{u}^i - z_i b &\geq 0, \\ \mathbf{u}^i &\geq 0 \end{aligned}$$



Knowledge-Based SVMs

The batch KBSVM formulation introduces ***advice slack variables*** to ***soften*** the advice constraints

$$
\begin{aligned}
\min \quad & \frac{1}{2}\|\mathbf{w}\|^{\mathbf{2}} + \lambda\,\mathbf{e}'\xi + \mu \sum_{\mathbf{i=1}}^{\mathbf{m}} \left(\eta^{\mathbf{i}} + \zeta_{\mathbf{i}}\right) \\
\text{s.t.} \quad & Y(X\mathbf{w} - b\mathbf{e}) + \xi \geq \mathbf{e}, \\
& \xi \geq 0, \\
& D_i'\mathbf{u}^i + z_i\mathbf{w} + \eta^i = 0, \\
& -\mathbf{d}^{i\prime}\mathbf{u}^i - z_i b + \zeta_i \geq 1, \\
& \mathbf{u}^i, \eta^i, \zeta_i \geq 0,\ i = 1, \ldots, m.
\end{aligned}
$$



Outline

- Knowledge-Based Support Vector Machines
- ***The Adviceptron: Online KBSVMs***
- A Real-World Task: Diabetes Diagnosis
- A Real-World Task: Tuberculosis Isolate Classification
- Conclusions

Online KBSVMs

- Need to derive an ***online version of KBSVMs***
- Algorithm is provided with advice and ***one*** labeled data point at each round
- Algorithm should ***update the hypothesis*** at each step, $\boldsymbol{w}^t$, ***as well as*** *the* ***advice vectors***, $\boldsymbol{u}^{i,t}$

Passive-Aggressive Algorithms

- Adopt the framework of ***passive-aggressive algorithms*** (Crammer et al, 2006), where at each round, when a new data point is given,
  - if loss = 0, there is no update (**passive**)
  - if loss > 0, update weights to minimize loss (**aggressive**)
- Why passive-aggressive algorithms?
  - readily applicable to **most SVM losses**
  - possible to derive elegant, **closed-form update rules**
  - simple rules provide fast updates; **scalable**
  - analyze performance by deriving **regret bounds**

Online KBSVMs

- There are *m* advice sets, $(D_i,\, \mathbf{d}^i,\, z_i)_{i=1}^m$
- At round *t*, the algorithm receives $(\mathbf{x}^t,\, y_t)$
- The current hypothesis is $\mathbf{w}^t$, and the current advice variables are $\mathbf{u}^{i,t},\, i = 1, \ldots, m$

At round *t*, the formulation for deriving an update is

$$\begin{array}{cl}
\min\limits_{\xi, \mathbf{u}^i, \eta^i, \zeta_i \geq 0, \mathbf{w}} & \frac{1}{2}\|\mathbf{w} - \mathbf{w}^t\|^2 + \frac{1}{2}\sum_{i=1}^m \|\mathbf{u}^i - \mathbf{u}^{i,t}\|^2 + \frac{\lambda}{2}\xi^2 + \frac{\mu}{2}\sum_{i=1}^m \left(\|\eta^i\|^2 + \zeta_i^2\right) \\
\text{subject to} & y_t \mathbf{w}'\mathbf{x}^t - 1 + \xi \geq 0, \\
& \left.\begin{array}{r} D_i'\mathbf{u}^i + z_i \mathbf{w} + \eta^i = 0 \\ -\mathbf{d}^{i'}\mathbf{u}^i - 1 + \zeta_i \geq 0 \\ \mathbf{u}^i \geq 0 \end{array}\right\} \; i = 1, \ldots, m.
\end{array}$$

Formulation At The *t*-th Round

- There are *m* advice sets, $(D_i,\ \mathbf{d}^i,\ z_i)_{i=1}^m$
- At round *t*, the algorithm receives $(\mathbf{x}^t,\ y_t)$
- The current hypothesis is $\mathbf{w}^t$, and the current advice variables are $\mathbf{u}^{i,t},\ i = 1, \ldots, m$

proximal terms for hypothesis and advice vectors

$$
\begin{array}{cl}
\displaystyle\min_{\xi, \mathbf{u}^i, \eta^i, \zeta_i \geq 0, \mathbf{w}} & \displaystyle \frac{1}{2}\|\mathbf{w} - \mathbf{w}^t\|^2 + \frac{1}{2}\sum_{i=1}^m \|\mathbf{u}^i - \mathbf{u}^{i,t}\|^2 + \frac{\lambda}{2}\xi^2 + \frac{\mu}{2}\sum_{i=1}^m \left(\|\eta^i\|^2 + \zeta_i^2\right), \\
\text{subject to} & y_t \mathbf{w}'\mathbf{x}^t - 1 + \xi \geq 0, \\
& \left.\begin{array}{r} D_i'\mathbf{u}^i + z_i\mathbf{w} + \eta^i = 0 \\ -\mathbf{d}^{i\prime}\mathbf{u}^i - 1 + \zeta_i \geq 0 \\ \mathbf{u}^i \geq 0 \end{array}\right\} \ i = 1, \ldots, m.
\end{array}
$$

Formulation At The *t*-th Round

- There are *m* advice sets, $(D_i,\, \mathbf{d}^i,\, z_i)_{i=1}^m$
- At round *t*, the algorithm receives $(\mathbf{x}^t,\, y_t)$
- The current hypothesis is $\mathbf{w}^t$, and the current advice variables are $\mathbf{u}^{i,t},\, i = 1, ..., m$

$$
\begin{array}{rl}
\min\limits_{\xi, \mathbf{u}^i, \eta^i, \zeta_i \geq 0, \mathbf{w}} & \frac{1}{2}\|\mathbf{w} - \mathbf{w}^t\|^2 + \frac{1}{2}\sum_{i=1}^m \|\mathbf{u}^i - \mathbf{u}^{i,t}\|^2 + \frac{\lambda}{2}\overset{\text{data loss}}{\xi^2} + \frac{\mu}{2}\sum_{i=1}^m \overset{\text{advice loss}}{\left(\|\eta^i\|^2 + {\zeta_i}^2\right)}, \\
\text{subject to} & y_t \mathbf{w}'\mathbf{x}^t - 1 + \xi \geq 0, \\
& \left.\begin{array}{r} D_i'\mathbf{u}^i + z_i\mathbf{w} + \eta^i = 0 \\ -\mathbf{d}^{i\prime}\mathbf{u}^i - 1 + \zeta_i \geq 0 \\ \mathbf{u}^i \geq 0 \end{array}\right\} \; i = 1, ..., m.
\end{array}
$$

Formulation At The *t*-th Round

- There are *m* advice sets, $(D_i,\ \mathbf{d}^i,\ z_i)_{i=1}^m$
- At round *t*, the algorithm receives $(\mathbf{x}^t,\ y_t)$
- The current hypothesis is $\mathbf{w}^t$, and the current advice variables are $\mathbf{u}^{i,t},\ i = 1, ..., m$

parameters

$$
\begin{aligned}
\min_{\xi, \mathbf{u}^i, \eta^i, \zeta_i \geq 0, \mathbf{w}} \quad & \frac{1}{2}\|\mathbf{w} - \mathbf{w}^t\|^2 + \frac{1}{2}\sum_{i=1}^m \|\mathbf{u}^i - \mathbf{u}^{i,t}\|^2 + \frac{\lambda}{2}\xi^2 + \frac{\mu}{2}\sum_{i=1}^m \left(\|\eta^i\|^2 + \zeta_i^2\right), \\
\text{subject to} \quad & y_t \mathbf{w}'\mathbf{x}^t - 1 + \xi \geq 0, \\
& \left.\begin{aligned} D_i'\mathbf{u}^i + z_i\mathbf{w} + \eta^i &= 0 \\ -\mathbf{d}^{i\prime}\mathbf{u}^i - 1 + \zeta_i &\geq 0 \\ \mathbf{u}^i &\geq 0 \end{aligned}\right\} \; i = 1, ..., m.
\end{aligned}
$$

Formulation At The *t*-th Round

- There are *m* advice sets, $(D_i,\, \mathbf{d}^i,\, z_i)_{i=1}^m$
- At round *t*, the algorithm receives $(\mathbf{x}^t,\, y_t)$
- The current hypothesis is $\mathbf{w}^t$, and the current advice variables are $\mathbf{u}^{i,t},\, i = 1, \ldots, m$

$$
\begin{array}{rl}
\displaystyle\min_{\xi,\mathbf{u}^i,\eta^i,\zeta_i \geq 0,\mathbf{w}} & \displaystyle\frac{1}{2}\|\mathbf{w}-\mathbf{w}^t\|^2 + \frac{1}{2}\sum_{i=1}^m \|\mathbf{u}^i - \mathbf{u}^{i,t}\|^2 + \frac{\lambda}{2}\xi^2 + \frac{\mu}{2}\sum_{i=1}^m \left(\|\eta^i\|^2 + \zeta_i^2\right), \\
\text{subject to} & y_t \mathbf{w}'\mathbf{x}^t - 1 + \xi \geq 0, \\
& \left.\begin{array}{r} D_i'\mathbf{u}^i + z_i \mathbf{w} + \eta^i = 0 \\ -\mathbf{d}^{i\prime}\mathbf{u}^i - 1 + \zeta_i \geq 0 \\ \mathbf{u}^i \geq 0 \end{array}\right\} i = 1, \ldots, m.
\end{array}
$$

inequality constraints make deriving a closed-form update impossible

Formulation At The *t*-th Round

- There are *m* advice sets, $(D_i, \mathbf{d}^i, z_i)_{i=1}^m$
- At round *t*, the algorithm receives $(\mathbf{x}^t, y_t)$
- The current hypothesis is $\mathbf{w}^t$, and the current advice-vector estimates are $\mathbf{u}^{i,t}$, $i = 1, \ldots, m$

DECOMPOSE INTO SMALLER SUBPROBLEMS

$$
\begin{aligned}
\min_{\xi, \mathbf{u}^i, \eta^i, \zeta_i \geq 0, \mathbf{w}} \quad & \frac{1}{2}\|\mathbf{w} - \mathbf{w}^t\|^2 + \frac{1}{2}\sum_{i=1}^{m}\|\mathbf{u}^i - \mathbf{u}^{i,t}\|^2 + \frac{\lambda}{2}\xi^2 + \frac{\mu}{2}\sum_{i=1}^{m}\left(\|\eta^i\|^2 + \zeta_i^2\right), \\
\text{subject to} \quad & y_t \mathbf{w}'\mathbf{x}^t - 1 + \xi \geq 0, \\
& \left.\begin{aligned} D_i'\mathbf{u}^i + z_i\mathbf{w} + \eta^i &= 0 \\ -\mathbf{d}^{i\prime}\mathbf{u}^i - 1 + \zeta_i &\geq 0 \\ \mathbf{u}^i &\geq 0 \end{aligned}\right\} \; i = 1, \ldots, m.
\end{aligned}
$$

inequality constraints make deriving a closed-form update impossible

Decompose Into *m+1* Sub-problems

- First sub-problem: update hypothesis by fixing the advice variables, to their values at the *t*-th iteration $\mathbf{u}^i = \mathbf{u}^{i,t}$

$$
\begin{aligned}
\min_{\xi, \mathbf{u}^i, \eta^i, \zeta_i \geq 0, \mathbf{w}} \quad & \frac{1}{2}\|\mathbf{w} - \mathbf{w}^t\|^2 + \frac{1}{2}\sum_{i=1}^{m}\|\mathbf{u}^i - \mathbf{u}^{i,t}\|^2 + \frac{\lambda}{2}\xi^2 + \frac{\mu}{2}\sum_{i=1}^{m}\left(\|\eta^i\|^2 + \zeta_i^2\right), \\
\text{subject to} \quad & y_t \mathbf{w}'\mathbf{x}^t - 1 + \xi \geq 0, \\
& \left.\begin{array}{r} D_i'\mathbf{u}^i + z_i\mathbf{w} + \eta^i = 0 \\ -\mathbf{d}^{i'}\mathbf{u}^i - 1 + \zeta_i \geq 0 \\ \mathbf{u}^i \geq 0 \end{array}\right\} \; i = 1, \ldots, m.
\end{aligned}
$$

- Some objective terms and constraints drop out of the formulation

Deriving The Hypothesis Update

- First sub-problem: update hypothesis by fixing the advice vectors

$$
\begin{aligned}
\mathbf{w}^{t+1} \;=\; \min_{\mathbf{w},\xi,\eta^i} \quad & \frac{1}{2}\|\mathbf{w}-\mathbf{w}^t\|_2^2 + \frac{\lambda}{2}\xi^2 + \frac{\mu}{2}\sum_{i=1}^{m}\|\eta^i\|_2^2 \\
\text{subject to} \quad & y_t\mathbf{w}'\mathbf{x}^t - 1 + \xi \geq 0, & (\alpha) \\
& D_i'\mathbf{u}^{i,t} + z_i\mathbf{w} + \eta^i = 0,\; i = 1,\ldots,m. & (\beta^i)
\end{aligned}
$$

Deriving The Hypothesis Update

- First sub-problem: update hypothesis by fixing the advice vectors

$$\begin{aligned} \mathbf{w}^{t+1} &= \min_{\mathbf{w},\xi,\eta^i} \quad \frac{1}{2}\|\mathbf{w}-\mathbf{w}^t\|_2^2 + \frac{\lambda}{2}\xi^2 + \frac{\mu}{2}\sum_{i=1}^{m}\|\eta^i\|_2^2 \\ &\text{subject to} \quad y_t\mathbf{w}'\mathbf{x}^t - 1 + \xi \geq 0, && (\alpha) \\ &\qquad\qquad D_i'\mathbf{u}^{i,t} + z_i\mathbf{w} + \eta^i = 0,\ i = 1,\ldots,m. && (\beta^i) \end{aligned}$$

fixed, ***advice-estimate of the hypothesis according to i-th advice set*** ; denote as $\mathbf{r}^{i,t}$

Advice-Estimate Of Current Hypothesis

- First sub-problem: update hypothesis by fixing the advice vectors

$$\mathbf{w}^{t+1} = \min_{\mathbf{w},\xi,\eta^i} \quad \frac{1}{2}\|\mathbf{w}-\mathbf{w}^t\|_2^2 + \frac{\lambda}{2}\xi^2 + \frac{\mu}{2}\sum_{i=1}^{m}\|\eta^i\|_2^2$$

$$\text{subject to} \quad y_t\mathbf{w}'\mathbf{x}^t - 1 + \xi \geq 0, \qquad (\alpha)$$

$$D_i'\mathbf{u}^{i,t} + z_i\mathbf{w} + \eta^i = 0, \; i = 1, \ldots, m. \qquad (\beta^i)$$

fixed, ***advice-estimate of the hypothesis according to i-th advice set*** ; denote as $\mathbf{r}^{i,t}$

average ***advice-estimates over all m advice vectors*** and denote as

$$\mathbf{r}^t = \frac{1}{m}\sum_{i=1}^{m}\mathbf{r}^{i,t}$$

The Hypothesis Update

For $\lambda,\ \mu > 0$, and advice-estimate $\mathbf{r}^t$, the hypothesis update is

$$\mathbf{w}^{t+1} \; = \; \nu\left(\mathbf{w}^t + \alpha_t y_t \mathbf{x}^t\right) + (1-\nu)\,\mathbf{r}^t,$$

$$\alpha_t = \left(\frac{1}{\lambda} + \nu\|\mathbf{x}^t\|^2\right)^{-1} \cdot \max\left(\, 1 - \nu\, y_t {\mathbf{w}^t}' \mathbf{x}^t - (1-\nu)\, y_t {\mathbf{r}^t}' \mathbf{x}^t \;,\; 0\right).$$

The Hypothesis Update

For λ, $\mu > 0$, and advice-estimate $\mathbf{r}^t$, the hypothesis update is

$$\mathbf{w}^{t+1} = \nu\left(\mathbf{w}^t + \alpha_t y_t \mathbf{x}^t\right) + (1-\nu)\,\mathbf{r}^t,$$

$$\alpha_t = \left(\frac{1}{\lambda} + \nu\|\mathbf{x}^t\|^2\right)^{-1} \cdot \max\left(1 - \nu\, y_t \mathbf{w}^{t\prime} \mathbf{x}^t - (1-\nu)\, y_t \mathbf{r}^{t\prime} \mathbf{x}^t\ ,\ 0\right).$$

Update is ***convex combination*** of the standard ***passive-aggressive update*** and the ***average advice-estimate***

Parameter of convex combinations is $\nu = \dfrac{1}{1+m\mu}$

The Hypothesis Update

For $\lambda,\ \mu > 0$, and advice-estimate $\mathbf{r}^t$, the hypothesis update is

$$\mathbf{w}^{t+1} \;=\; \nu\left(\mathbf{w}^t + \alpha_t y_t \mathbf{x}^t\right) + (1-\nu)\,\mathbf{r}^t,$$

$$\alpha_t = \left(\frac{1}{\lambda} + \nu\|\mathbf{x}^t\|^2\right)^{-1} \cdot \max\left(1 - \nu\, y_t \mathbf{w}^{t\prime}\mathbf{x}^t - (1-\nu)\, y_t \mathbf{r}^{t\prime}\mathbf{x}^t\ ,\ 0\right).$$

Update is ***convex combination*** of the standard ***passive-aggressive update*** and the ***average advice-estimate***

Update weight depends on ***hinge loss*** computed with respect to a ***composite weight vector*** that is a ***convex combination*** of the ***current hypothesis*** and the ***average advice-estimate***

Parameter of convex combinations is $\nu = \dfrac{1}{1+m\mu}$

Deriving The Advice Updates

- Second sub-problem: update advice vectors by fixing the hypothesis $\mathbf{w} = \mathbf{w}^{t+1}$

$$
\begin{aligned}
\min_{\xi, \mathbf{u}^i, \eta^i, \zeta_i \geq 0, \mathbf{w}} \quad & \frac{1}{2}\|\mathbf{w} - \mathbf{w}^t\|^2 + \frac{1}{2}\sum_{i=1}^{m} \|\mathbf{u}^i - \mathbf{u}^{i,t}\|^2 + \frac{\lambda}{2}\xi^2 + \frac{\mu}{2}\sum_{i=1}^{m}\left(\|\eta^i\|^2 + \zeta_i^2\right), \\
\text{subject to} \quad & y_t \mathbf{w}'\mathbf{x}^t - 1 + \xi \geq 0, \\
& \left.\begin{aligned}
D_i'\mathbf{u}^i + z_i \mathbf{w} + \eta^i &= 0 \\
-\mathbf{d}^{i\prime}\mathbf{u}^i - 1 + \zeta_i &\geq 0 \\
\mathbf{u}^i &\geq 0
\end{aligned}\right\} \; i = 1, \ldots, m.
\end{aligned}
$$

- Some constraints and objective terms drop out of the formulation

Deriving The Advice Updates

- Second sub-problem: update advice vectors by fixing the hypothesis

$$
\begin{array}{rl}
\displaystyle\min_{\mathbf{u}^i,\eta^i,\zeta_i\geq 0,} & \displaystyle\frac{1}{2}\sum_{i=1}^{m}\|\mathbf{u}^i-\mathbf{u}^{i,t}\|^2+\frac{\mu}{2}\sum_{i=1}^{m}\left(\|\eta^i\|^2+\zeta_i^2\right), \\
\text{subject to} & \left.\begin{aligned} D_i'\mathbf{u}^i+z_i\mathbf{w}^{t+1}+\eta^i&=0\\ -\mathbf{d}^{i'}\mathbf{u}^i-1+\zeta_i&\geq 0\\ \mathbf{u}^i&\geq 0\end{aligned}\right\}\ i=1,\ldots,m.
\end{array}
$$

Deriving The Advice Updates

- Second sub-problem: update advice vectors by fixing the hypothesis

$$
\begin{array}{rl}
\displaystyle \min_{\mathbf{u}^i,\eta^i,\zeta_i \geq 0,} & \displaystyle \frac{1}{2}\sum_{i=1}^{m} \|\mathbf{u}^i - \mathbf{u}^{i,t}\|^2 + \frac{\mu}{2}\sum_{i=1}^{m} \left(\|\eta^i\|^2 + \zeta_i^2\right), \\
\text{subject to} & \left.\begin{array}{r} D_i'\mathbf{u}^i + z_i\mathbf{w}^{t+1} + \eta^i = 0 \\ -\mathbf{d}^{i\prime}\mathbf{u}^i - 1 + \zeta_i \geq 0 \\ \mathbf{u}^i \geq 0 \end{array}\right\} \; i = 1, \ldots, m.
\end{array}
$$

split into *m* sub-problems

Deriving The i-th Advice Updates

- *m* sub-problems: update the *i*-th advice vector by fixing the hypothesis

$$
\begin{aligned}
\mathbf{u}^{i,t+1} = \min_{\mathbf{u}^i,\eta,\zeta} \quad & \frac{1}{2}\|\mathbf{u}^i - \mathbf{u}^{i,t}\|^2 + \frac{\mu}{2}\left(\|\eta^i\|_2^2 + \zeta_i^2\right) \\
\text{subject to} \quad & D_i'\mathbf{u}^i + z_i\mathbf{w}^{t+1} + \eta^i = 0, && (\beta^i) \\
& -\mathbf{d}^{i\prime}\mathbf{u}^i - 1 + \zeta_i \geq 0, && (\gamma_i) \\
& \mathbf{u}^i \geq 0. && (\tau^i)
\end{aligned}
$$

Deriving The i-th Advice Updates

- *m* sub-problems: update the *i*-th advice vector by fixing the hypothesis

$$
\begin{aligned}
\mathbf{u}^{i,t+1} = \min_{\mathbf{u}^i,\eta,\zeta} \quad & \frac{1}{2}\|\mathbf{u}^i - \mathbf{u}^{i,t}\|^2 + \frac{\mu}{2}\left(\|\eta^i\|_2^2 + \zeta_i^2\right) \\
\text{subject to} \quad & D_i'\mathbf{u}^i + z_i\mathbf{w}^{t+1} + \eta^i = 0, && (\beta^i) \\
& -\mathbf{d}^{i\prime}\mathbf{u}^i - 1 + \zeta_i \geq 0, && (\gamma_i) \\
& \mathbf{u}^i \geq 0. && (\tau^i)
\end{aligned}
$$

cone constraints still complicating

cannot derive closed form solution

- ***Use projected-gradient approach***
  - ***drop constraints*** to compute intermediate closed-form update
  - ***project*** intermediate update back on to cone constraints

The *m* Advice Updates

For $\mu > 0$, and current hypothesis $\mathbf{w}^{t+1}$, for each advice set, $i = 1, \ldots, m$, the update rule is given by

$$\mathbf{u}^{i,t+1} \;=\; \max\left(\mathbf{u}^{i,t} + D_i\beta^i - \mathbf{d}^i\,\gamma_i,\; 0\right),$$

$$\begin{bmatrix} \beta^i \\ \gamma_i \end{bmatrix} = \begin{bmatrix} -(D_i'D_i + \frac{1}{\mu}I_n) & D_i'\mathbf{d}^i \\ \mathbf{d}^{i\prime}D_i & -(\mathbf{d}^{i\prime}\mathbf{d}^i + \frac{1}{\mu}) \end{bmatrix}^{-1} \begin{bmatrix} D_i'\mathbf{u}^{i,t} + z_i\mathbf{w}^{t+1} \\ -\mathbf{d}^{i\prime}\mathbf{u}^{i,t} - 1 \end{bmatrix}$$

The *m* Advice Updates

For $\mu > 0$, and current hypothesis $\mathbf{w}^{t+1}$, for each advice set, $i = 1, \ldots, m$, the update rule is given by

$$\mathbf{u}^{i,t+1} = \max\left(\mathbf{u}^{i,t} + D_i\beta^i - \mathbf{d}^i\,\gamma_i,\ 0\right),$$

projection

$$\begin{bmatrix} \beta^i \\ \gamma_i \end{bmatrix} = \begin{bmatrix} -(D_i'D_i + \frac{1}{\mu}I_n) & D_i'\mathbf{d}^i \\ \mathbf{d}^{i\prime}D_i & -(\mathbf{d}^{i\prime}\mathbf{d}^i + \frac{1}{\mu}) \end{bmatrix}^{-1} \begin{bmatrix} D_i'\mathbf{u}^{i,t} + z_i\mathbf{w}^{t+1} \\ -\mathbf{d}^{i\prime}\mathbf{u}^{i,t} - 1 \end{bmatrix}$$

- ***hypothesis-estimate of the advice;*** denote $\mathbf{s}^i = \beta^i/\gamma_i$
- The update is the ***error*** or the amount of ***violation of the constraint*** $D_i\mathbf{x} \leq \mathbf{d}^i$ by an ideal data point, $\mathbf{s}^i$

each advice update ***depends on the newly updated hypothesis***

The Adviceptron

input: data $(\mathbf{x}^t, y_t)_{t=1}^T$, advice sets $(D_i, \mathbf{d}^i, z_i)_{i=1}^m$, parameters λ, $\mu > 0$

initialize: $\mathbf{u}^{i,1} = 0$, $\mathbf{w}^1 = 0$

let: $\nu = 1/(1 + m\mu)$

for $(\mathbf{x}^t,\, y_t)$ **do**

predict label $\hat{y}_t = \mathrm{sign}(\mathbf{w}^{t\prime}\mathbf{x}^t)$

receive correct label y_t

suffer loss

$$\ell_t = \max\left(1 - \nu y_t \mathbf{w}^{t\prime}\mathbf{x}^t - (1-\nu) y_t \mathbf{r}^{t\prime}\mathbf{x}^t,\ 0\right)$$

update hypothesis using $\mathbf{u}^{i,t}$

$$\alpha = \ell_t / (\frac{1}{\lambda} + \nu \|\mathbf{x}^t\|^2), \quad \mathbf{w}^{t+1} = \nu\,(\,\mathbf{w}^t + \alpha\, y_t \mathbf{x}^t\,) + (1-\nu)\,\mathbf{r}^t$$

update advice variables using $\mathbf{w}^{t+1}$

$$(\beta^i,\, \gamma_i) = H_i^{-1}\mathbf{g}^i, \quad \mathbf{u}^{i,t+1} = \left(\mathbf{u}^{i,t} + D_i\beta^i - \mathbf{d}^i\gamma_i\right)_+$$

end for

Outline

- Knowledge-Based Support Vector Machines
- The Adviceptron: Online KBSVMs
- ***A Real-World Task: Diabetes Diagnosis***
- A Real-World Task: Tuberculosis Isolate Classification
- Conclusions

Diagnosing Diabetes

- Standard data set from UCI repository (768 x 8)
  - all patients at least 21 years old of Pima Indian heritage
  - features include body mass index, blood glucose level
- Expert advice for diagnosing diabetes from NIH website on risks for Type-2 diabetes
  - a person who is obese (characterized by BMI > 30) and has a high blood glucose level (> 126) is at a strong risk for diabetes

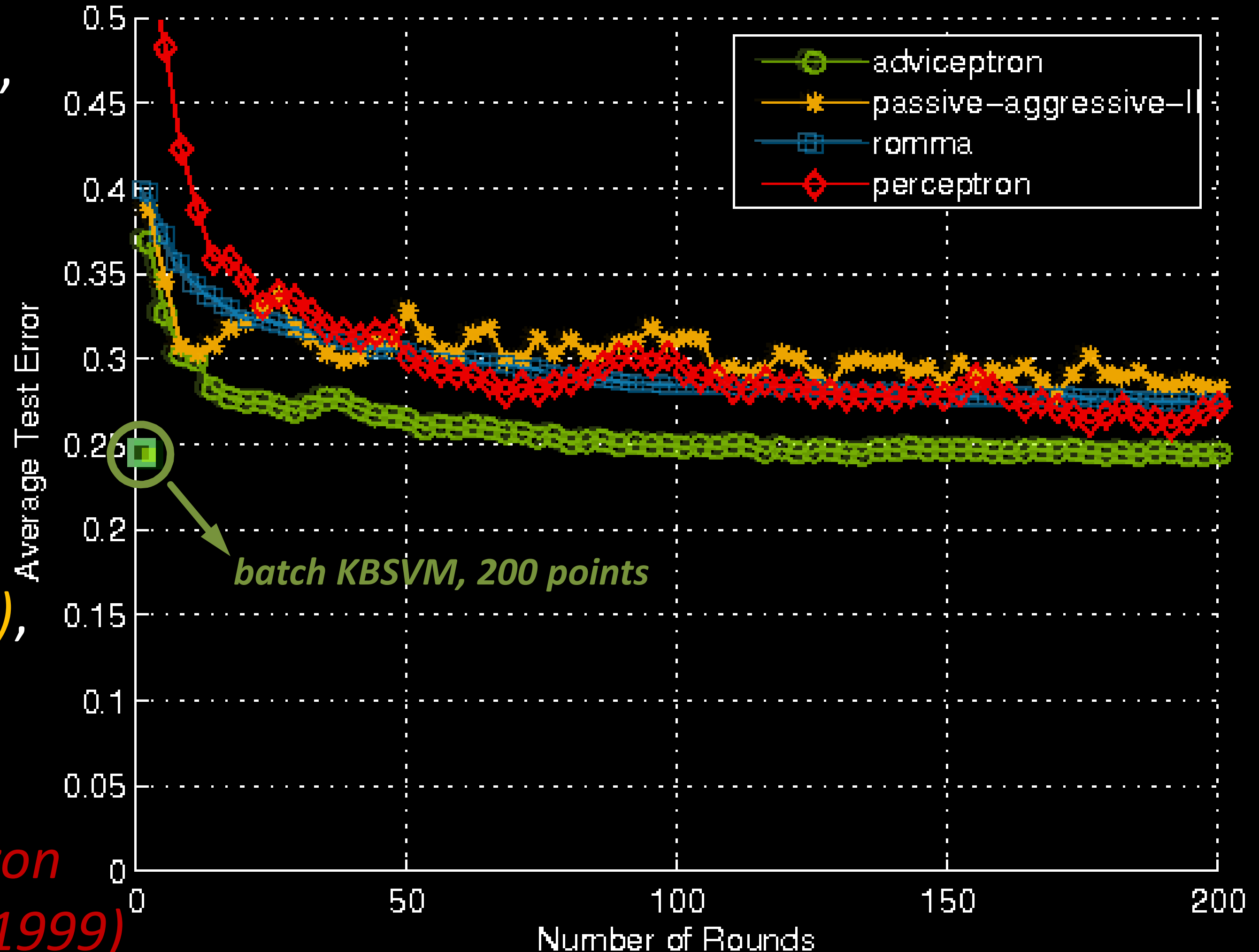
$$(\mathtt{BMI} \geq 30) \wedge (\mathtt{bloodglucose} \geq 126) \Rightarrow \mathtt{diabetes}$$

  - a person who is at normal weight (BMI < 25) and has low blood glucose level (< 100) is at a low risk for diabetes

$$(\mathtt{BMI} \leq 25) \wedge (\mathtt{bloodglucose} \leq 100) \Rightarrow \neg\mathtt{diabetes}$$

Diagnosing Diabetes: Results

- 200 examples for training, remaining for testing
- Results averaged over 20 randomized iterations
- Compared to advice-free online algorithms:
  - *Passive-aggressive (Crammer et al, 2006)*,
  - *ROMMA (Li & Long, 2002)*,
  - *Max margin-perceptron (Freund & Schapire, 1999)*



Outline

- Knowledge-Based Support Vector Machines
- The Adviceptron: Online KBSVMs
- A Real-World Task: Diabetes Diagnosis
- ***A Real-World Task: Tuberculosis Isolate Classification***
- Conclusions

Tuberculosis Isolate Classification

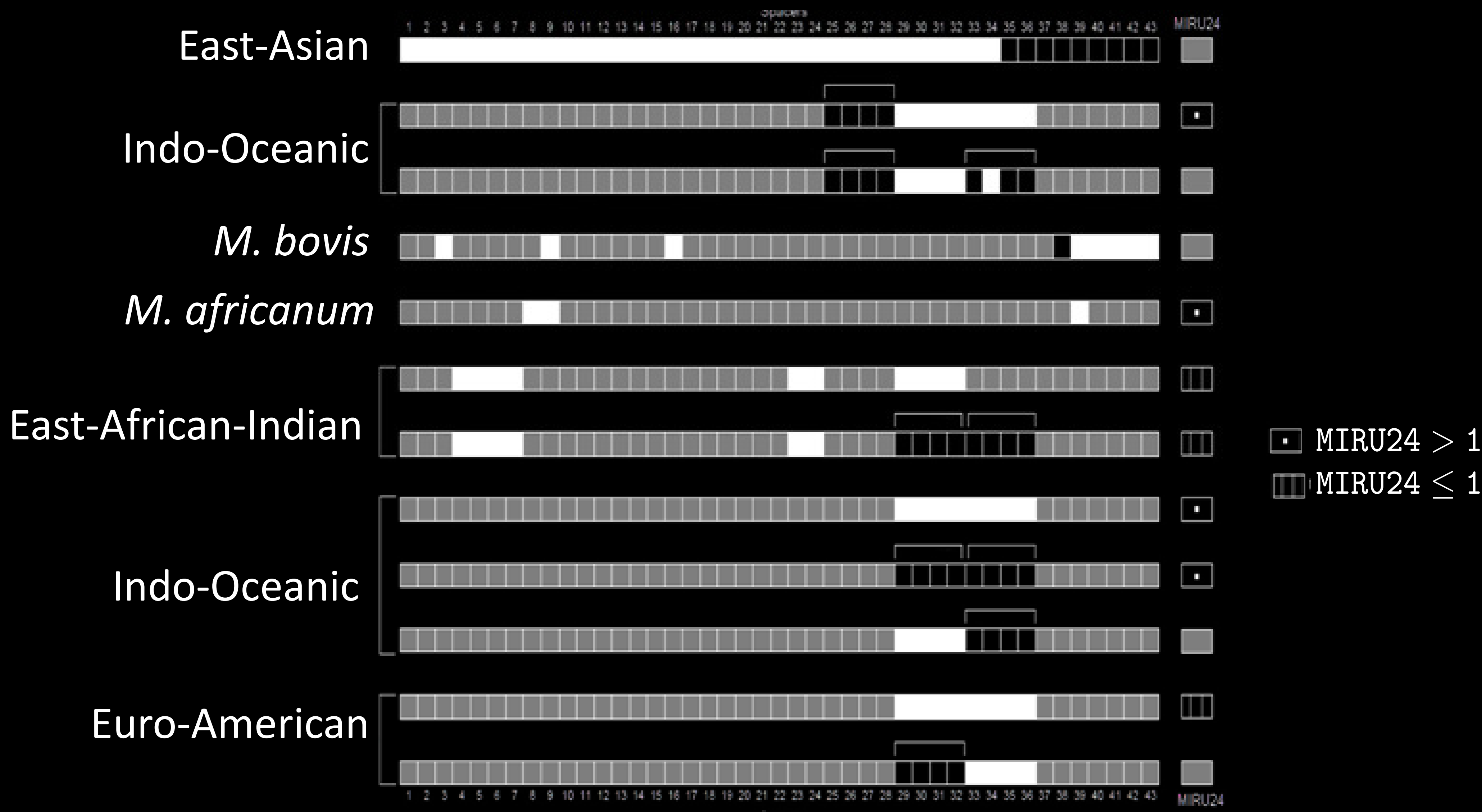
- Task is to classify strains of *Mycobacterium tuberculosis* complex (MTBC) into major genetic lineages based on DNA fingerprints
- MTBC is the causative agent for TB
  - leading cause of disease and morbidity
  - strains vary in infectivity, transmission, virulence, immunogenicity, host associations depending on genetic lineage
- Lineage classification is crucial for surveillance, tracking and control of TB world-wide

Tuberculosis Isolate Classification

- Two types of DNA fingerprints for all culture-positive TB strains collected in the US by the CDC (44 data features)
- Six (classes) major lineages of TB for classification
  - ***ancestral:*** *M. bovis, M. africanum,* Indo-Oceanic
  - ***modern:*** Euro-American, East-Asian, East-African-Indian
- Problem formulated as six 1-vs-many classification tasks

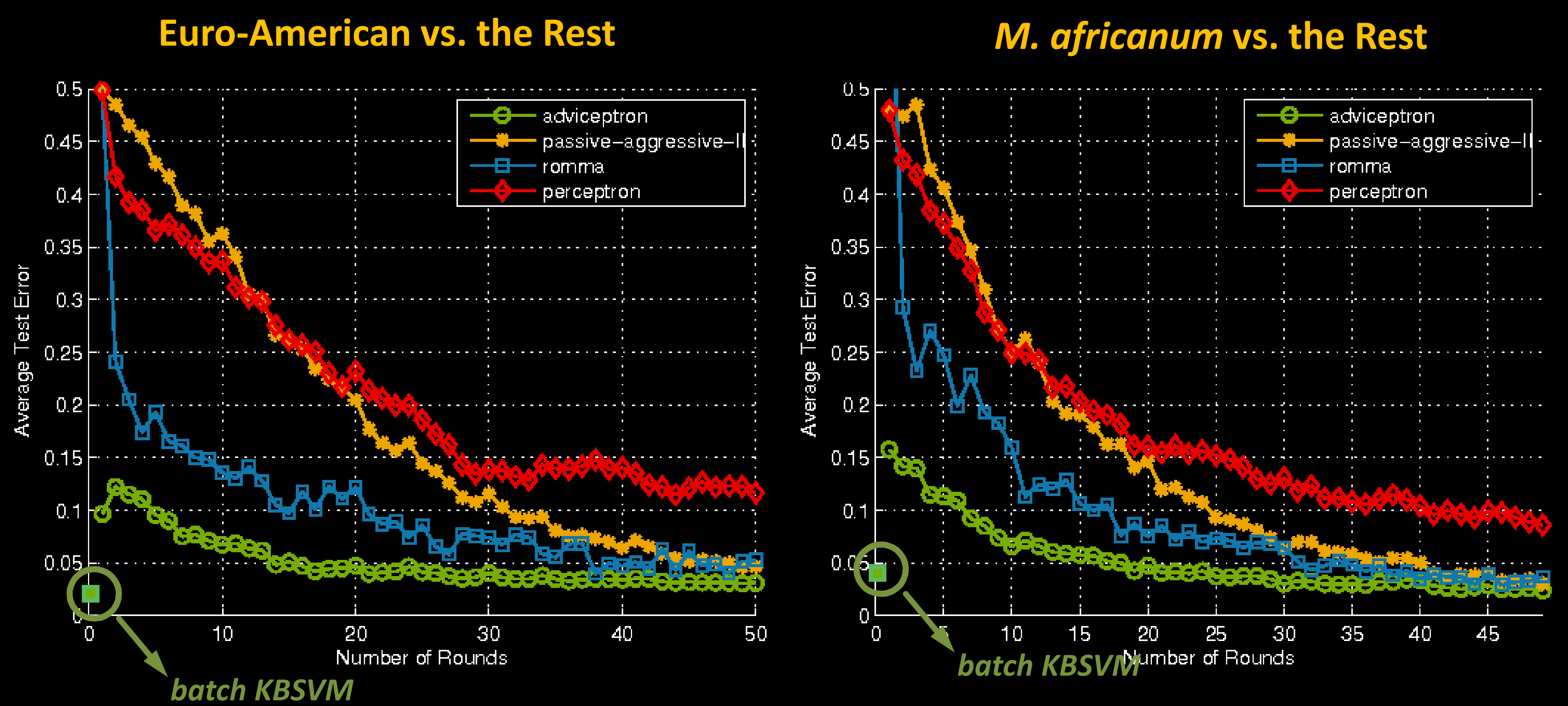
Class	#isolates	#pieces of Positive Advice	#pieces of Negative Advice
East-Asian	4924	1	1
East-African-Indian	1469	2	4
Euro-American	25161	1	2
Indo-Oceanic	5309	5	5
M. africanum	154	1	3
M. bovis	693	1	3

Expert Rules for TB Lineage Classification

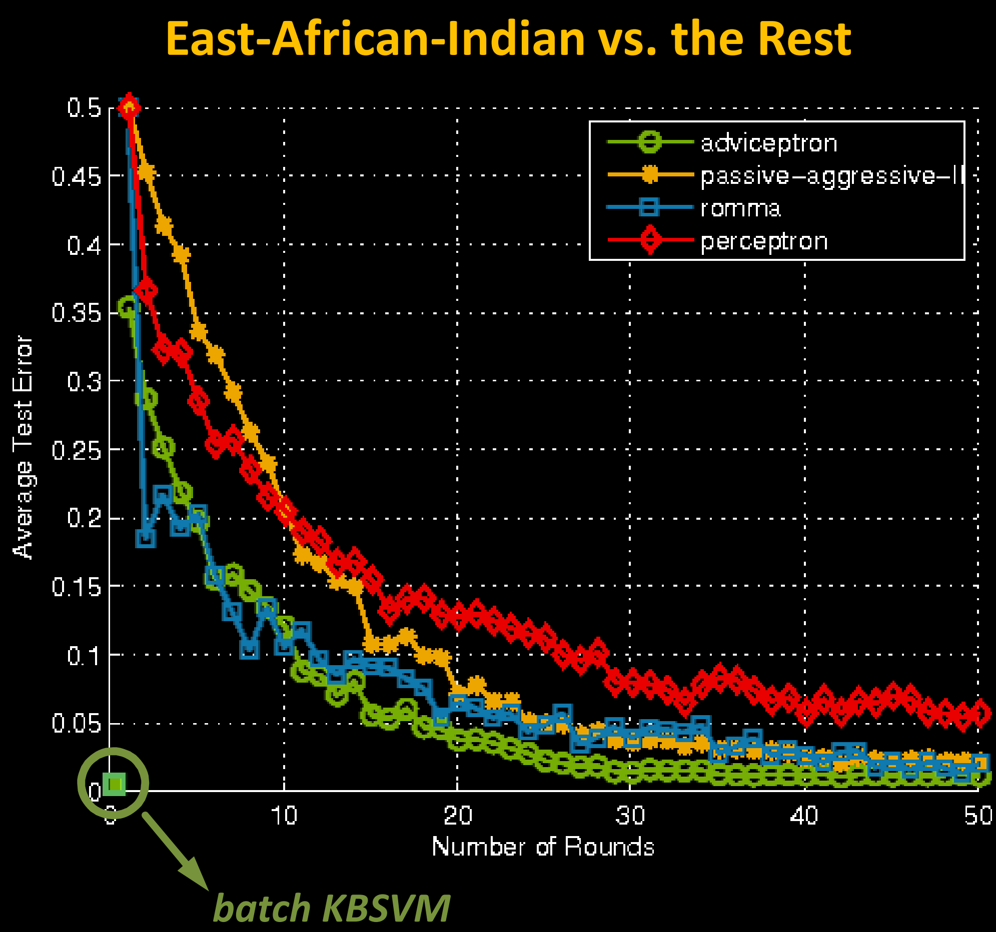


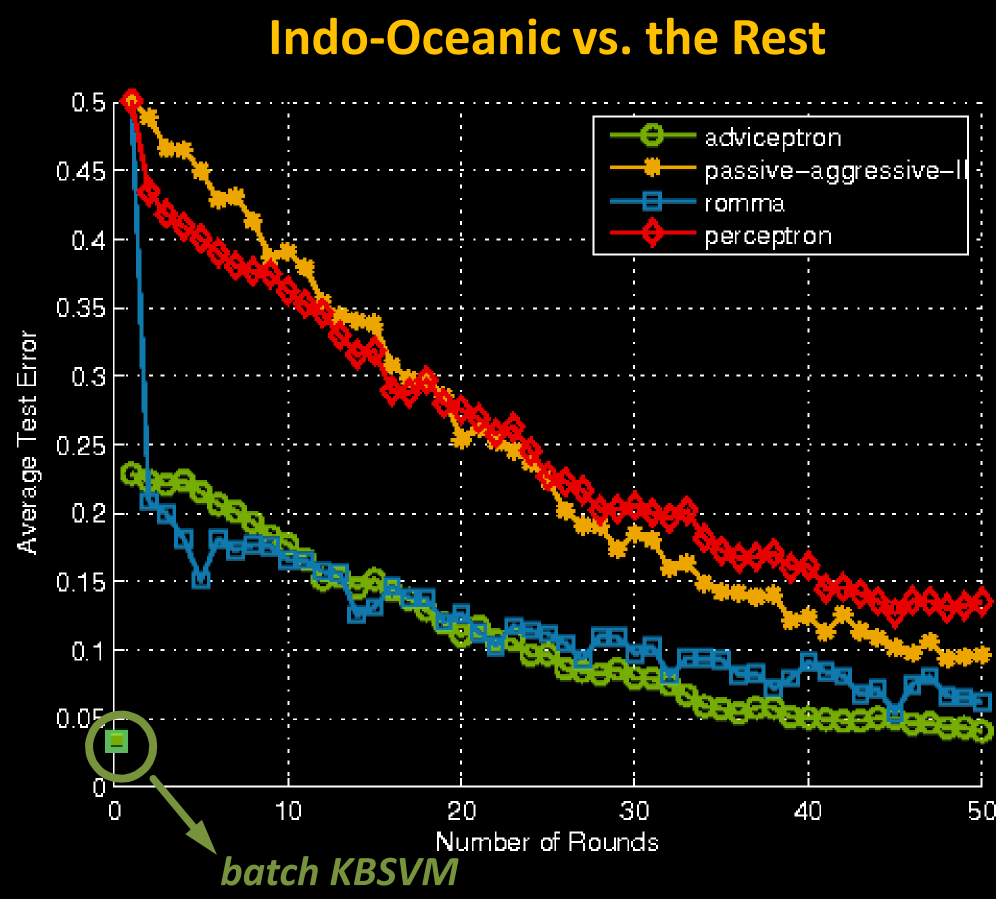
Rules provided by Dr. Lauren Cowan at the Center for Disease Control, documented in Shabbeer et al, (2010)

TB Results: Might Need Fewer Examples To Converge With Advice



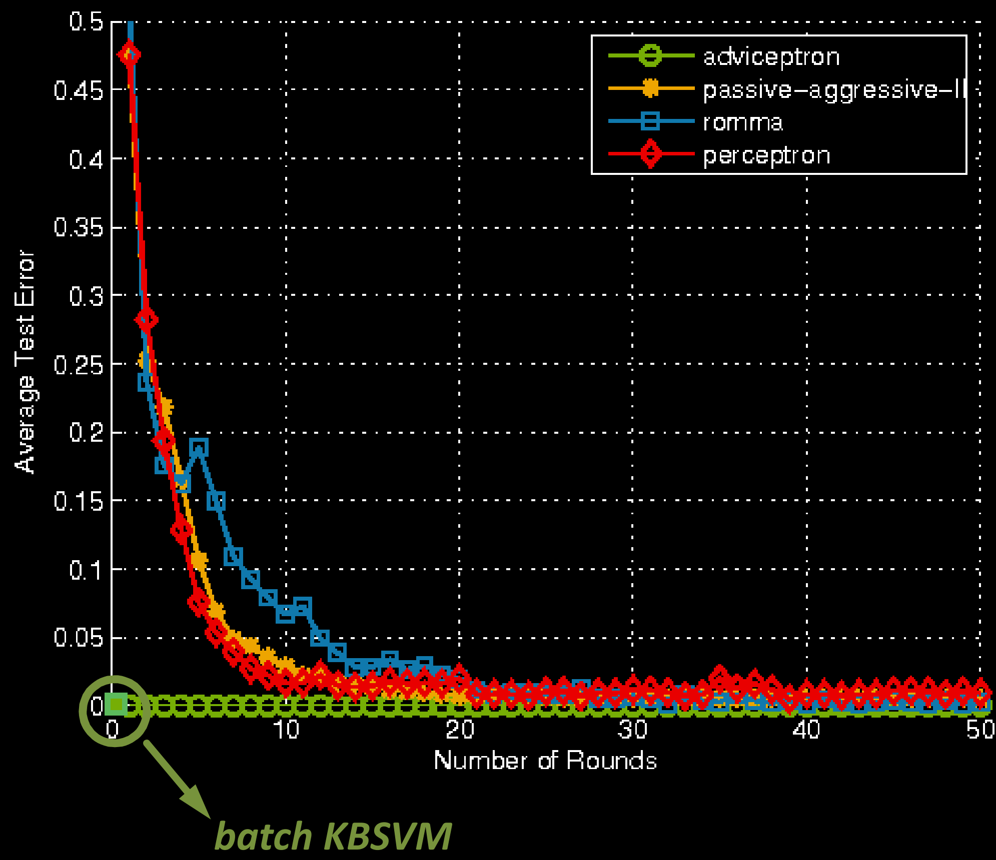
TB Results: Can Converge To A Better Solution With Advice



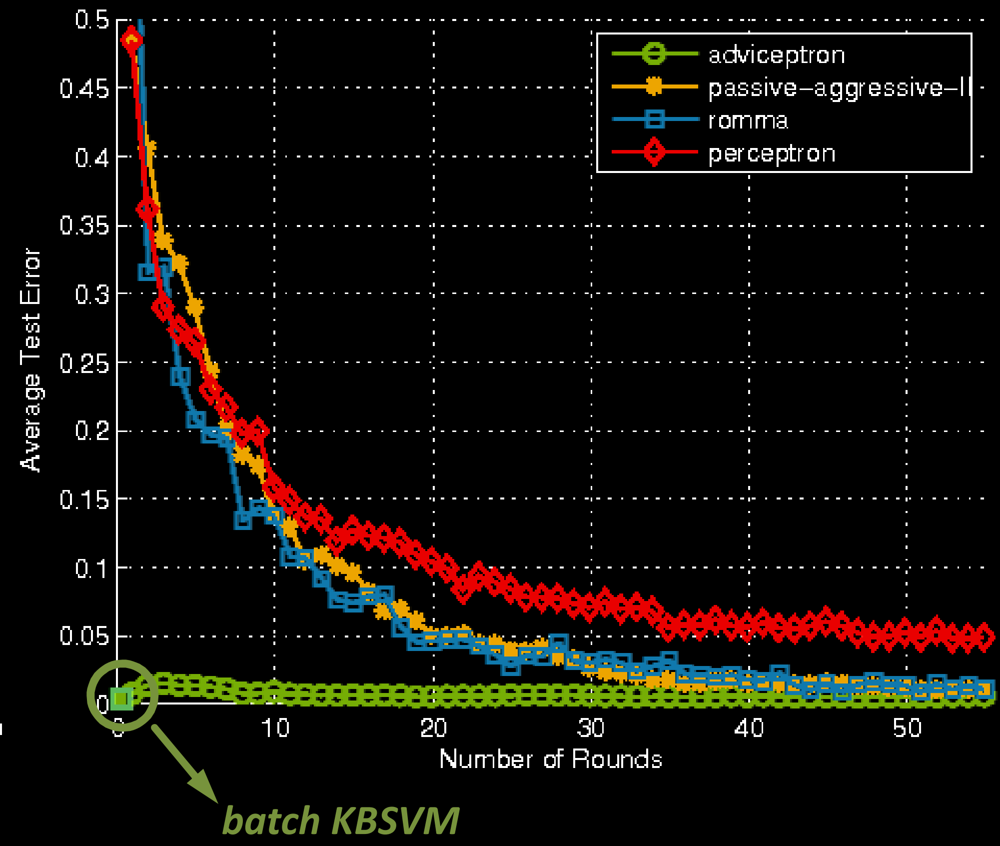


TB Results: Possible To Still Learn Well With ***Only*** Advice

East-Asian vs. the Rest



M. bovis **vs. the Rest**



Outline

- Knowledge-Based Support Vector Machines
- The Adviceptron: Online KBSVMs
- A Real-World Task: Diabetes Diagnosis
- A Real-World Task: Tuberculosis Isolate Classification
- ***Conclusions And Questions***

Conclusions

- New online learning algorithm: ***the adviceptron***
- Makes use of prior knowledge in the form of (possibly imperfect) ***polyhedral advice***
- Performs ***simple, closed-form updates*** via passive-aggressive framework; scalable
- Good advice can help converge to a ***better solution with fewer examples***
- ***Encouraging empirical results*** on two important real-world tasks

References.

(Fung et al, 2003) G. Fung, O. L. Mangasarian, and J. W. Shavlik. *Knowledge-based support vector classifiers*. In S. Becker, S. Thrun & K. Obermayer, eds, NIPS, 15, pp. 521–528, 2003

(Crammer et al, 2006) K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. *Online passive-aggressive algorithms*. J. of Mach. Learn. Res., 7:551–585, 2006.

(Freund and Schapire, 1999) Y. Freund and R. E. Schapire. *Large margin classification using the perceptron algorithm*. Mach. Learn., 37(3):277–296, 1999.

(Li and Long, 2002) Y. Li and P. M. Long. *The relaxed online maximum margin algorithm*. Mach. Learn., 46(1/3):361–387, 2002.

(Shabbeer et al, 2001) A. Shabbeer, L. Cowan, J. R. Driscoll, C. Ozcaglar, S. L Vandenberg, B. Yener, and K. P Bennett. *TB-Lineage: An online tool for classification and analysis of strains of* Mycobacterium tuberculosis *Complex*. Unpublished manuscript, 2010.

Acknowledgements.

The authors would like to thank Dr. Lauren Cowan of the Center for Disease Control (CDC) for providing the TB dataset and the expert-defined rules for lineage classification.

We gratefully acknowledge support of DARPA under grant **HR0011-07-C-0060** *and the NIH under grant* **1-R01-LM009731-01**.

Views and conclusions contained in this document are those of the authors and do not necessarily represent the official opinion or policies, either expressed or implied of the US government or of DARPA.

KBSVMs: Deriving The Advice Constraints

We assume an expert provides polyhedral advice of the form

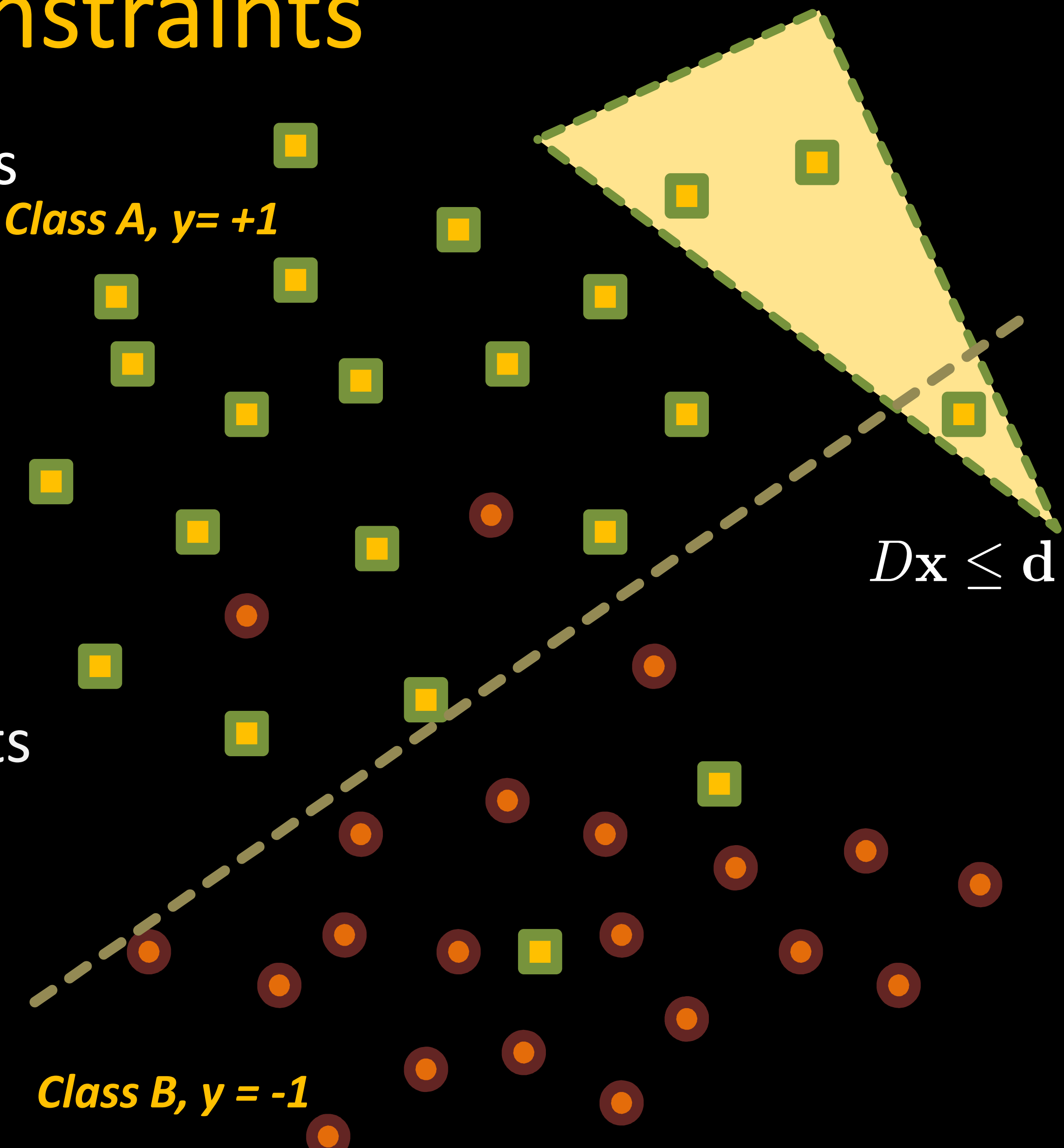
$$D\mathbf{x} \leq \mathbf{d} \Rightarrow \mathbf{w}'\mathbf{x} \geq b$$

We know $p \Rightarrow q$ is equivalent to $\neg p \vee q$

If $\neg p \vee q$ has a solution then its **negation has no solution** or,

$$\begin{aligned} D\mathbf{x} - \mathbf{d}\,\tau &\leq 0, \\ \mathbf{w}'\mathbf{x} - b\,\tau &< 0, \\ -\tau &< 0 \end{aligned}$$

has no solution $(\mathbf{x}, \tau)$.



KBSVMs: Deriving The Advice Constraints

If the following system

$$\begin{aligned} D\mathbf{x} - \mathbf{d}\,\tau &\leq 0, \\ \mathbf{w}'\mathbf{x} - b\,\tau &< 0, \\ -\tau &< 0 \end{aligned}$$

has no solution $(\mathbf{x}, \tau)$, then by ***Motzkin's Theorem of the Alternative***, the following system

$$\begin{aligned} D'\mathbf{u} + \mathbf{w} &= 0, \\ -\mathbf{d}'\mathbf{u} - b &\geq 0, \\ \mathbf{u} &\geq 0 \end{aligned}$$

has a solution u.