On the Consistency of Multithreshold Entropy Linear Classifier

Wojciech M. Czarnecki

19 February 2015

tfml 2015
theoretical foundations
of machine learning, Będlewo

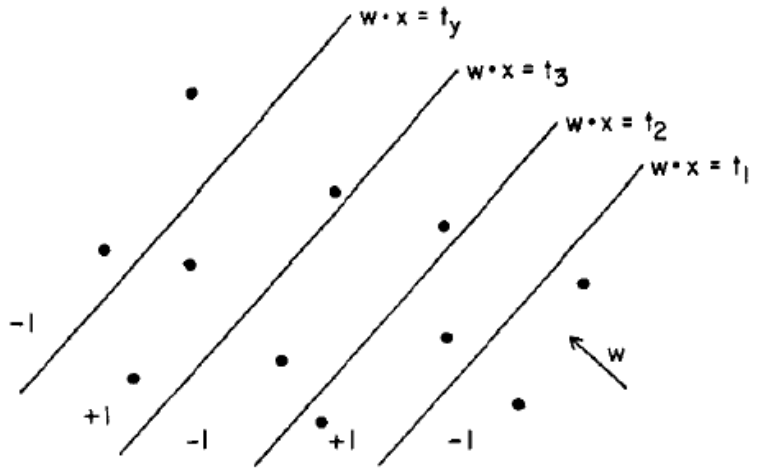
Linear classification with one and multiple thresholds

Linear classifier

$$cl(x;\ w, t) = \text{sign}(\langle w, x\rangle + t)$$

k-threshold Linear Classifier

$$cl(x;\ w, \{t_i\}_{i=1}^{k}) = \prod_{i=1}^{k} \text{sign}(\langle w, x\rangle + t_i)$$



Note

It is highly nontrivial to find a good multithreshold linear classifier. It is not clear whether efficient algorithm for finding even 2-threshold linear classifier exists, nor what is the exact Vapnik-Chervonenkis dimension of such family of hypotheses.

Idea

Idea (Multithreshold Entropy Linear Classifier)

Let us look for such a linear projection w for which that distributions of projected classes are as divergent as possible.

How we can quantiatively express how "good" is projection w?

- project each class from the training set on w
- perform kernel density estimation of this one-dimensional data (**why is it so important?**)
- compute Cauchy-Schwarz Divergence between estimated densities
- classify data using simple density-based classification

Definition (*W.M. Czarnecki and J. Tabor, 2014*)

Multithreshold Entropy Linear Classifier is the density classifier based on the one-dimensional projection of data on the $w \in \mathbb{R}^d$ such that it maximizes $\mathrm{D}_{CS}([\![w^T X_-]\!], [\![w^T X_+]\!])$.

$[\![X]\!]$ is the kernel density estimation of X. Optimization of such an objective function can be done (there exist analytical forms of all required equations), however it is quite expensive.

$$\begin{aligned}\text{maximize}_{w\in\mathbb{R}^d}\ \ & \mathrm{D}_{CS}([\![w^T X_-]\!],[\![w^T X_+]\!]) \\ & = 2H_2^{\times}([\![w^T X_-]\!],[\![w^T X_+]\!]) \\ & \quad - H_2([\![w^T X_-]\!]) - H_2([\![w^T X_+]\!])\end{aligned}$$

- $H_2(f) = -\log \int f^2(x)dx$
- $H_2^{\times}(f,g) = -\log \int fg(x)dx$

- H_2 are **regularization** terms (Renyi's quadratic entropy)
- $H_2^{\times}$ are **fitting** terms (Renyi's quadratic cross entropy)

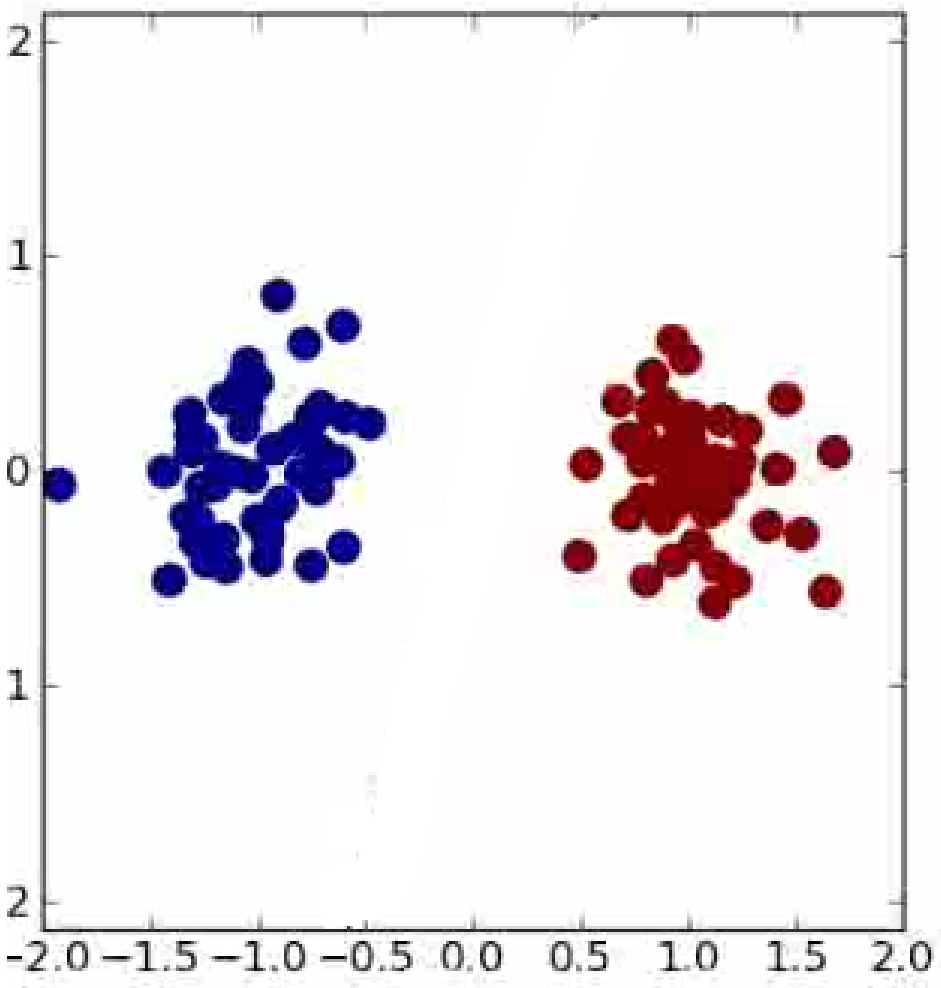
- Regularized MELC

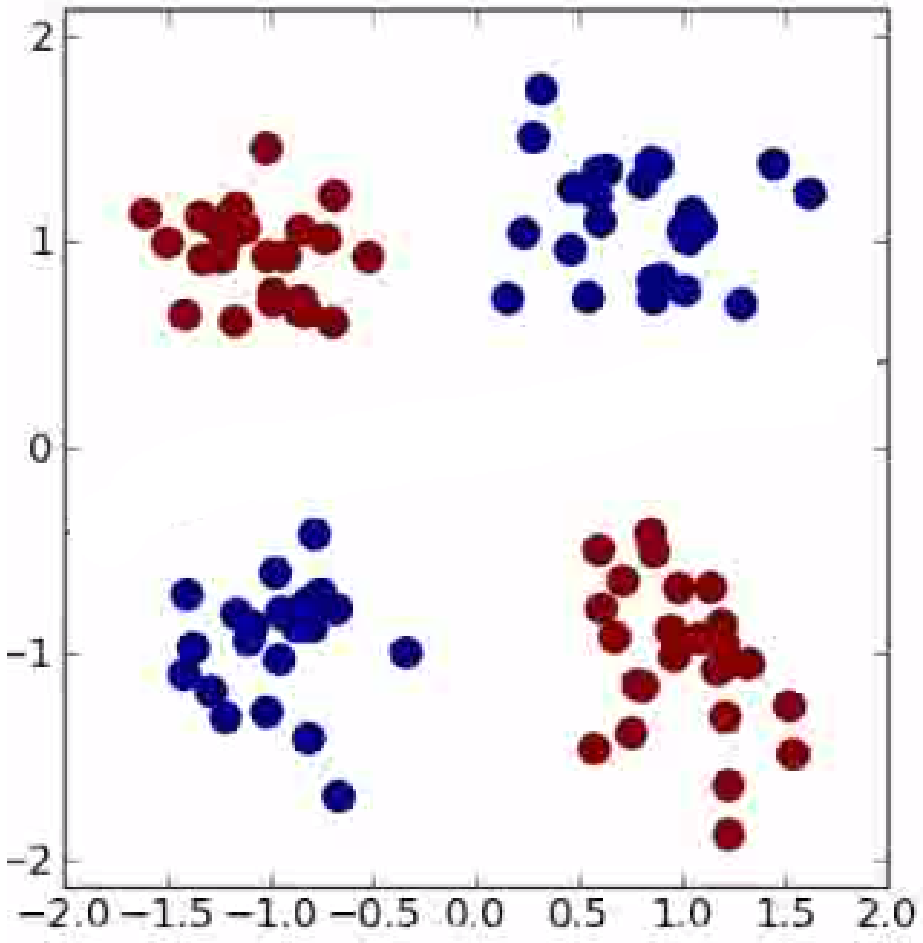
$$\text{maximize}_{w \in \mathbb{R}^d} \quad -2 \log \int fg(x)dx + \log \int f^2(x)dx + \log \int g^2(x)dx$$

- Non-regularized MELC

$$\text{maximize}_{w \in \mathbb{R}^d} \quad -2 \log \int fg(x)dx$$

for $f = w^T F, g = w^T G$


2
1
0
1
2
-2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0


2
1
0
−1
−2
−2.0 −1.5 −1.0 −0.5 0.0 0.5 1.0 1.5 2.0

Consistency

In very simple terms the consistency of the machine learning method is its ability to approximate any data probability distribution with **smallest possible classification error under some evaluation metric** given enough training points.

In other words we would expect that given infinitely many training points our method converges to smallest obtainable error in terms of some interesting metric.

In classification, one of the most basic and important evaluation metrics is the accuracy, which is directly connected to minimizing the sum of $0/1$ loss functions values

$$\mathrm{acc}(y, p) = 1 - \frac{1}{N} \sum_{i=1}^{N} \ell_{0/1}(y_i, p_i)$$

$$\ell_{0/1}(y, p) = 1 \iff py \geq 0$$

Most of the existing machine learning algorithms change this loss function (which is hard to optimize, non-continuous, lacks much information) to some **upper bound**, for example hinge loss used in SVM because **its optimization is tracktable, easier, faster**

$$\ell_H(y, p) = \max\{0, 1 - py\}$$

so $\ell_H(y, p) \geq \ell_{0/1}(y, p)$ and $\ell_H(y, p) = 0 \rightarrow \ell_{0/1}(y, p) = 0$

Due to the simple, additive nature of both optimization criterion and the evaluation metric it is rather easy to show its consistency, but what about **models that do not optimize any additive point-based loss function**?

Definition (Expected averaged accuracy)

Given a probability distributions f_-, f_+ the expected averaged accuracy of classifier cl is

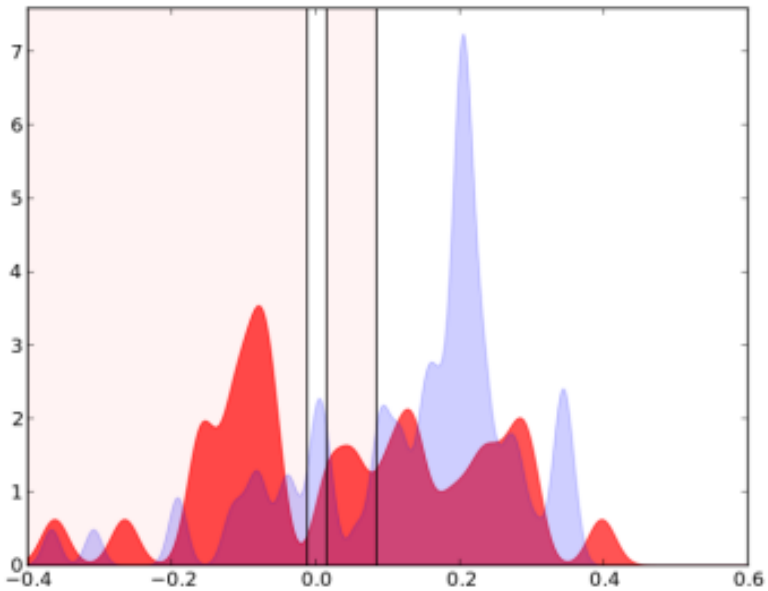
$$\frac{1}{2}\int \max\{0, -cl(x)\} f_-(x)dx + \frac{1}{2}\int \max\{0, cl(x)\} f_+(x)dx$$

In other words EAA expresses **probability of misclassification** done by our model cl assuming that classes are equaly probable.

Observation

For the family of multithreshold linear classifiers, the smallest obtainable averaged classification error for f_-, f_+ is

$$\min_w \int \min\{(w^T f_-)(x), (w^T f_+)(x)\} dx$$



Note

For simplicity, we assume in futher slides that we are working on the continuous data distributions f_-, f_+ instead of the samples X_-, X_+ as all results we are interested in hold in the limiting case when the sample size grows to infinity, so kernel density estimation is arbitrary close to the true distributions

Simple distributions' families

- **goal:** $\int \min\{f(x), g(x)\}dx$
- **criterion:** $-2\log\int(fg)(x)dx$

Observation

Considered model is consistent with **multithreshold linearly separable distributions**.

Observation

Considered model is consistent with **radial normal distributions**.

It seems that it is not possible to prove the general consistency with given optimization criterion, however we can show analogous of loss function upper bounding, but on the level of whole expected errors.

Theorem

Negative log-likelihood of minimal misclassification error of a given multithreshold linear classifier for any non multithreshold linearly separable distributions is at least a half of Renyi's quadratic cross entropy of data projections used by this classifier.

Sketch of the proof

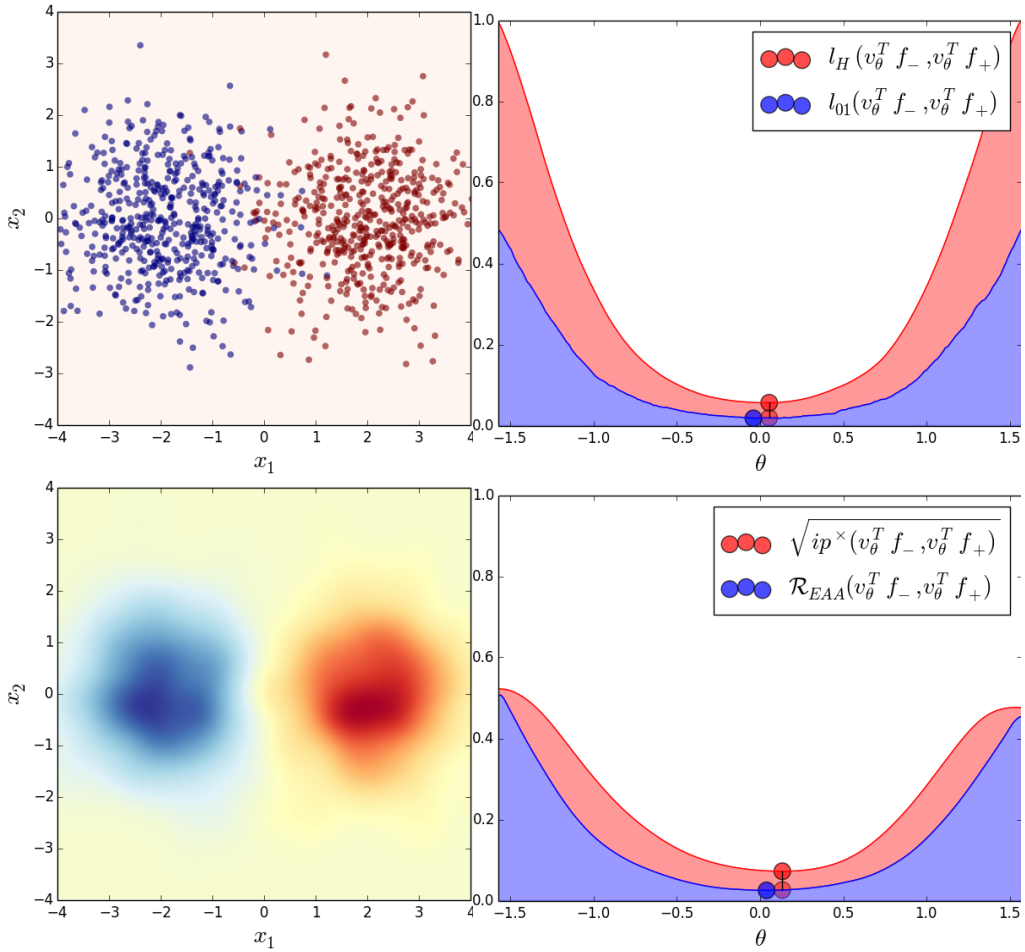
Using Schwarz inequality and $\min\{a, b\} \leq \sqrt{ab}$ we get that

$$\mathcal{R} = \int_0^1 \min\{w^T f_-(x), w^T f_+(x)\} dx \leq \sqrt{\int_0^1 w^T f_-(x) w^T f_+(x)}$$

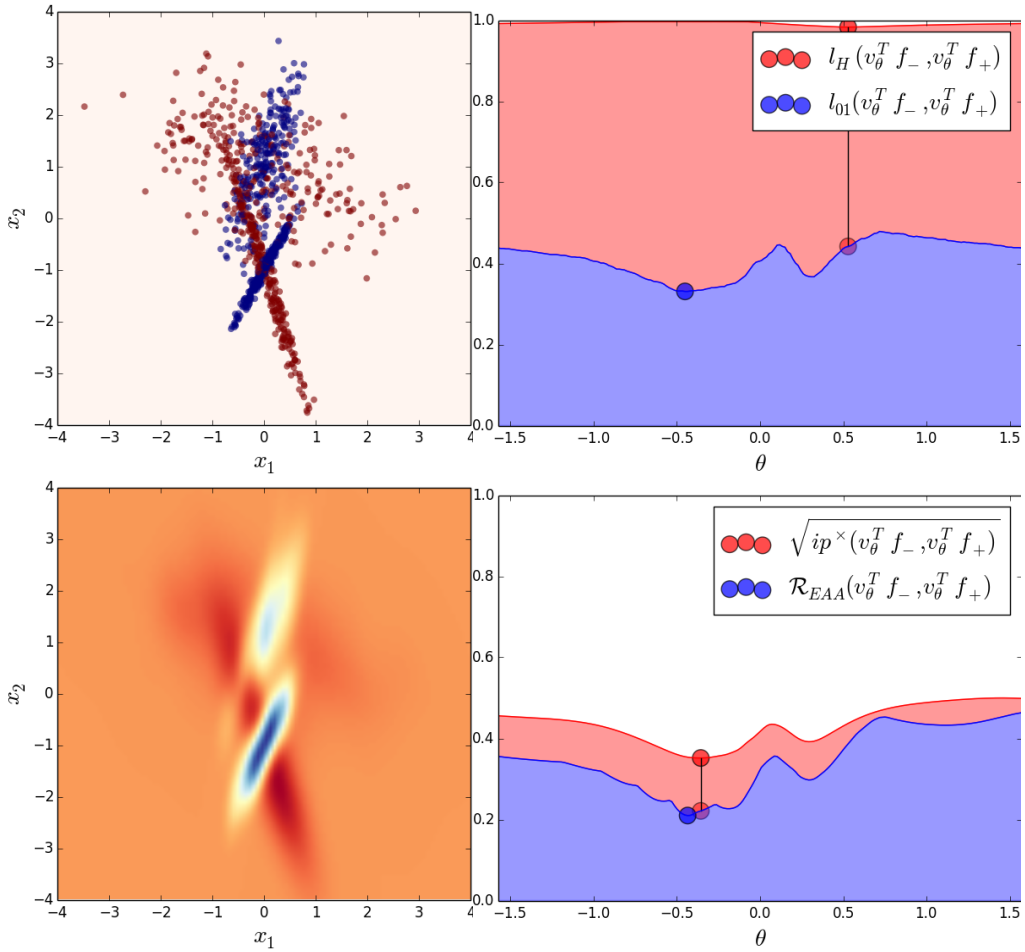
and as they are not separable, $\mathcal{R}$ is positive so

$$-\ln(\mathcal{R}) \geq -\ln\left(\sqrt{\int_0^1 w^T f_-(x) w^T f_+(x)}\right) = \tfrac{1}{2} H_2^{\times}(w^T f_-(x), w^T f_+(x))$$

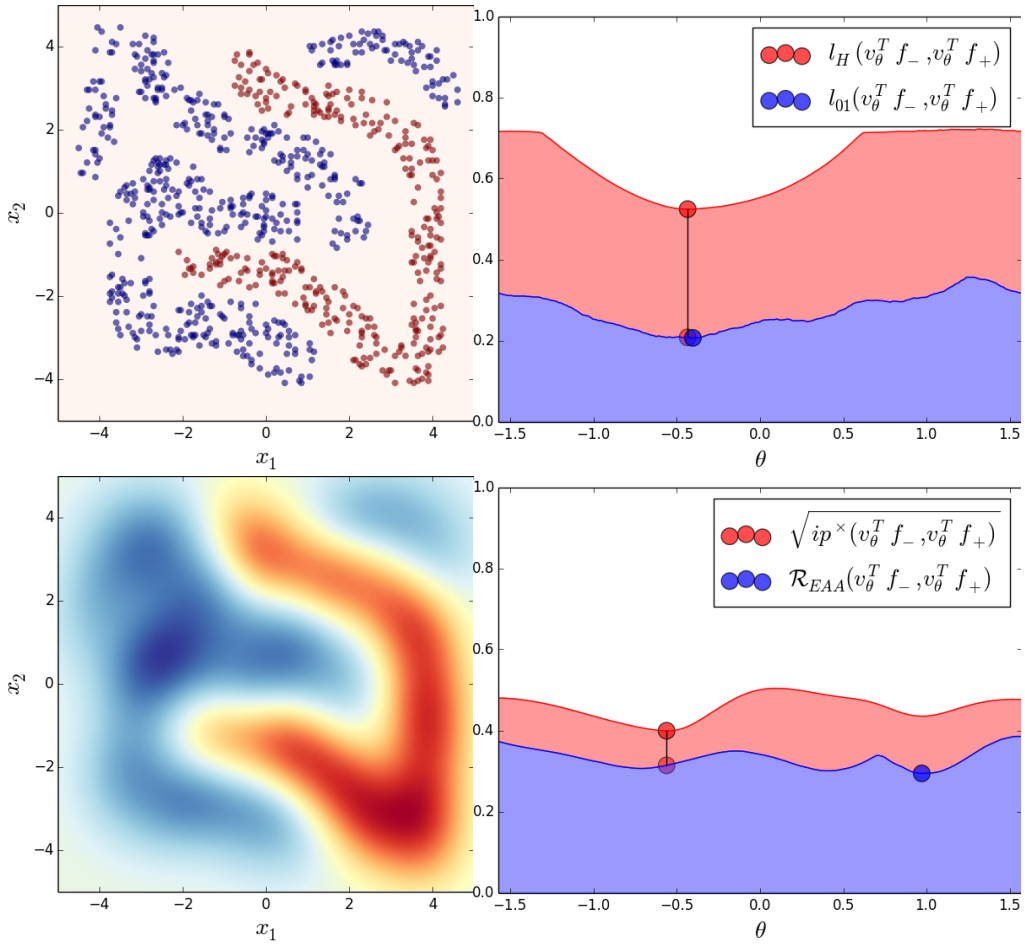
Examples


$l_H(v_\theta^T f_-, v_\theta^T f_+)$
$l_{01}(v_\theta^T f_-, v_\theta^T f_+)$
$\sqrt{ip^\times(v_\theta^T f_-, v_\theta^T f_+)}$
$\mathcal{R}_{EAA}(v_\theta^T f_-, v_\theta^T f_+)$
x_1
x_2
θ

Examples


$l_H(v_\theta^T f_-, v_\theta^T f_+)$
$l_{01}(v_\theta^T f_-, v_\theta^T f_+)$
$\sqrt{ip^\times(v_\theta^T f_-, v_\theta^T f_+)}$
$\mathcal{R}_{EAA}(v_\theta^T f_-, v_\theta^T f_+)$
x_1
x_2
θ

Examples


$l_H\,(v_\theta^T\,f_-\,,v_\theta^T\,f_+)$
$l_{01}(v_\theta^T\,f_-\,,v_\theta^T\,f_+)$
$\sqrt{ip^\times(v_\theta^T\,f_-\,,v_\theta^T\,f_+)}$
$\mathcal{R}_{EAA}(v_\theta^T\,f_-\,,v_\theta^T\,f_+)$
x_1
x_2
θ

Summary

- MELC is consistent with some simple distributions families
- MELC bounds the true averaged misclassification probability in the similar fashion hinge loss bounds the missclassification error
- It appears that even though it is not convex, it nicely smooths out the error surface making learning procedure more tracktable