THE UNIVERSITY OF TOKYO

ICML2011読む会

局所線形SVM

“*Locally Linear Support Vector Machines*”

L‘ubor Ladicky & Philip H.S. Torr (オックスフォード大の皆様)

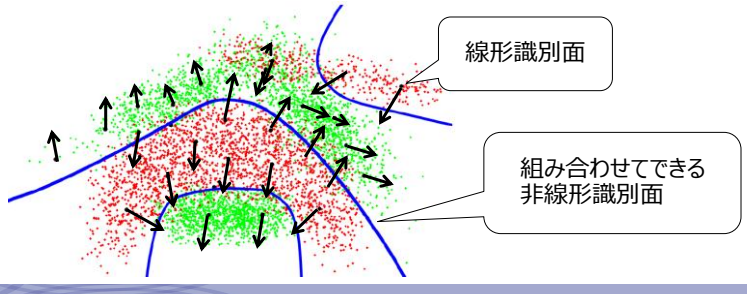
読むひと：鹿島 久嗣 東京大学 THE UNIVERSITY OF TOKYO

DEPARTMENT OF MATHEMATICAL INFORMATICS

The figures and tables are taken from the original paper.

この論文は
学習・予測を効率化するために 局所線形な識別器を提案しています

- 高速な非線形識別器をつくるのが目的
- 複数のアンカー点においた線形モデルを組み合わせることによって非線形識別器を構成する
  - 局所線形埋込み（LLE; Locally Linear Embedding）と似た考え方
- 最適化問題はSVMのそれと似た2次計画問題になる
  - 確率的勾配法で解く



モチベーション：
高速な非線形識別器を構成したい

- SVMなどの識別器はカーネル化することによって非線形な識別を行うことができる
- カーネル法は計算時間がデータ数に大きく依存するため、一般的に計算時間がかかる
- 一方、（計算時間が次元に大きく依存する）線形モデルに限った場合には効率的なインプリが出そろいつつある
- 非線形の場合にも効率的な識別器がほしい

モデル：
入力依存パラメータをもつ 局所線形な識別器を考える

- 線形モデルの形をもち、パラメータ$\mathbf{w}$が入力に依存して変わるようなモデルを考える

$$H(\mathbf{x}) = \mathbf{w}(\mathbf{x})^\top \mathbf{x} + b(\mathbf{x})$$

入力ベクトル

バイアス項

パラメータ $\mathbf{w}$ が入力 $\mathbf{x}$ に依存

- これだけでは何も言っていないのに等しいので、$\mathbf{w}(\mathbf{x})$ や $b(\mathbf{x})$ に何らかの制約を入れる必要がある

アイディア：
入力依存パラメータを少数のアンカー点におけるパラメータで近似する

- 入力依存パラメータを少数のアンカーで近似

$$\mathbf{w}(\mathbf{x}) = \sum_{\mathbf{v}\in\mathcal{C}} \gamma_{\mathbf{v}}(\mathbf{x})\, \mathbf{w}_{\mathbf{v}}$$

点$\mathbf{v}$におけるパラメータ

  - 少数のデータ点集合 C における線形モデルで全体を近似
  - $\gamma_{\mathbf{v}}(\mathbf{x})$ は線形モデルを組み合わせる係数
    - $\mathbf{v}$と$\mathbf{x}$間の距離について減少する関数などにする（学習しない）
- 線形モデルの形をもち、パラメータ$\mathbf{w}$ が入力に依存して変わるようなモデルを考える

$$H(\mathbf{x}) = \sum_{\mathbf{v}\in\mathcal{C}} \gamma_{\mathbf{v}}(\mathbf{x})\, \mathbf{w}_{\mathbf{v}}{}^{\top}\mathbf{x} + \gamma_{\mathbf{v}}(\mathbf{x})\, b_{\mathbf{v}}$$

$$= \boldsymbol{\gamma}(\mathbf{x})^{\top}\mathbf{W}\,\mathbf{x} + \boldsymbol{\gamma}(\mathbf{x})^{\top}\mathbf{b}$$

行列パラメータ

最適化問題：
提案モデルでふつうにフィット

- ふつうのSVMと同様、各事例に対するヒンジ損失＋L2正則化で目的関数を定義：

$$\arg\min_{\mathbf{W},\mathbf{b}} \frac{\lambda}{2}||\mathbf{W}||^2 + \frac{1}{|S|}\sum_{k\in S}\max(0, 1 - y_k H_{\mathbf{W},\mathbf{b}}(\mathbf{x}_k)), \quad (9)$$

  - もしくは：

$$\begin{aligned} &\arg\ \min_{\mathbf{W},\mathbf{b}} && \frac{\lambda}{2}||\mathbf{W}||^2 + \frac{1}{|S|}\sum_{k\in S}\xi_k \qquad (10)\\ &s.t. \forall k \in S: && \xi_k \geq 0 \\ & && \xi_k \geq 1 - y_k(\gamma(\mathbf{x}_k)^T\mathbf{W}\mathbf{x}_k + \gamma(\mathbf{x}_k)^T\mathbf{b}). \end{aligned}$$

アルゴリズム：ふつうに確率的勾配法

- データをひとつ（$\mathbf{x}_t$）もってきて、損失関数の勾配方向にパラメータ更新

$$\mathbf{W}_{t+1} = \mathbf{W}_t + \frac{1}{\lambda(t+t_0)} y_t(\mathbf{x}_t \gamma(\mathbf{x}_t)^T) \quad (11)$$

$$\mathbf{b}_{t+1} = \mathbf{b}_t + \frac{1}{\lambda(t+t_0)} y_t \gamma(\mathbf{x}_t), \quad (12)$$

t_0:適当な定数

- （効率化のため）正則化は何ステップかに一度おこなう

正則化の周期

$$\mathbf{W}'_{t+1} = \mathbf{W}_{t+1}(1 - \frac{skip}{t+t_0}). \quad (13)$$

実験：数字認識・文字認識タスクにおいて予測精度、速度について既存手法と比較

- 数字認識・文字認識タスクで評価（MNIST, USPS, LETTER）
- たとえばMNIST（40000データ）の場合：
  - 100個のアンカー点をk-meansクラスタリングで求める
  - $\gamma_v(\mathbf{x})$は8近傍を用い、近傍への距離について減少する重み
- 予測は（全データ数に比較して少ない）アンカー点のなかから、近いものを（この場合8個）とってくるだけなので高速

アンカー点の数と予測精度の関係の検証：
数字認識タスクにおいて、アンカー点は100点程度で十分

- MNIST（40,000データ）の場合：

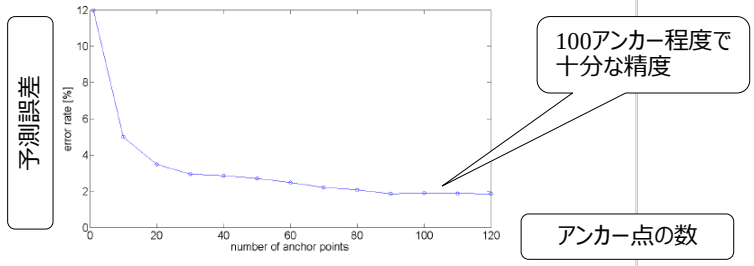


Figure 2. Dependency of the performance of LL-SVM *on number of anchor points on* MNIST *data set. Standard linear* SVM *is equivalent to the* LL-SVM *with one anchor point. The performance is saturated at around* 100 *anchor points due to insufficiently large amount of training data.*

従来手法との比較：
提案法は学習・予測ともに高速

- MNIST（40,000データ）の場合：

Table 1. A comparison of the performance, training and test times of LL-SVM *with the state-of the art algorithms on* MNIST *data set. All kernel* SVM *methods (Chang & Lin, 2001; Bordes et al., 2005; Crammer & Singer, 2002; Tsochantaridis et al., 2005; Bordes et al., 2007) used* RBF *kernel. Our method achieved comparable performance to the state-of-the-art and could be seen as a good trade-off between very fast linear* SVM *and the qualitatively best kernel methods.* LL-SVM *was approximately* 50-3000 *times faster than different kernel based methods. As the complexity of kernel methods grow more than linearly, we expected larger relative difference for larger data sets. Running times of* MCSVM, SVM_{struct} *and* LA-RANK *are as reported in (Bordes et al., 2007) and thus only illustrative. N/A means the running times are not available.)*

Method	error	training time	test time
Linear SVM (Bordes et al., 2009) (10 passes)	12.00%	1.5 s	8.75 μs
Linear SVM on LCC (Yu et al., 2009) (512 a.p.)	2.64%	N/A	N/A
Linear SVM on LCC (Yu et al., 2009) (4096 a.p.)	1.90%	N/A	N/A
LibSVM (Chang & Lin, 2001)	1.36%	17500 s	46 ms
LA-SVM (Bordes et al., 2005) (1 pass)	1.42%	4900 s	40.6 ms
LA-SVM (Bordes et al., 2005) (2 passes)	1.36%	12200 s	42.8 ms
MCSVM (Crammer & Singer, 2002)	1.44%	25000 s	N/A
SVM_{struct} (Tsochantaridis et al., 2005)	1.40%	265000 s	N/A
LA-RANK (Bordes et al., 2007) (1 pass)	1.41%	30000 s	N/A
LL-SVM (100 a.p., 10 passes)	1.85%	81.7 s	470 μs

この論文は
学習・予測を効率化するために 局所線形な識別器を提案しています

- 複数のアンカー点においた線形モデルを組み合わせることによって高速な非線形識別器を構成した
- 簡単で そこそこ うまくいくので、悪くはないのでは？
- せっかく線形モデルが得られるのだから、解釈性の向上という方向もあるのではないか？
- テクニカルには 重み $\gamma_v(\mathbf{x})$ を固定しない方法が美しいだろう
  → fused lasso的アプローチ？