Support Vector Machine with Pinball Loss and its Algorithms

Xiaolin Huang, Lei Shi, Johan A.K. Suykens

KU Leuven, ESAT-STADIUS

BIL2014
15th, July, 2014

Outline

1. SVM with pinball loss
   - quantile distance
   - pin-SVM
   - properties
2. Solving algorithms
   - primal algorithm
   - sequential minimal optimization
   - solution path
3. Numerical example
4. Conclusion

Outline

1. SVM with pinball loss
   - quantile distance
   - pin-SVM
   - properties
2. Solving algorithms
   - primal algorithm
   - sequential minimal optimization
   - solution path
3. Numerical example
4. Conclusion

Outline

1. SVM with pinball loss
   - quantile distance
   - pin-SVM
   - properties
2. Solving algorithms
   - primal algorithm
   - sequential minimal optimization
   - solution path
3. Numerical example
4. Conclusion

minimal distance

- let $\mathbf{z} = \{x_i, y_i\}_{i=1}^m$, where $x_i \in \mathbb{R}^n$ and $y_i \in \{-1, 1\}$. Then $\mathbf{z}$ consists of two classes with the following sets of indices: $\mathbf{I} = \{i \mid y_i = 1\}$ and $\mathbf{II} = \{i \mid y_i = -1\}$
- SVM tries to find a hyperplane via

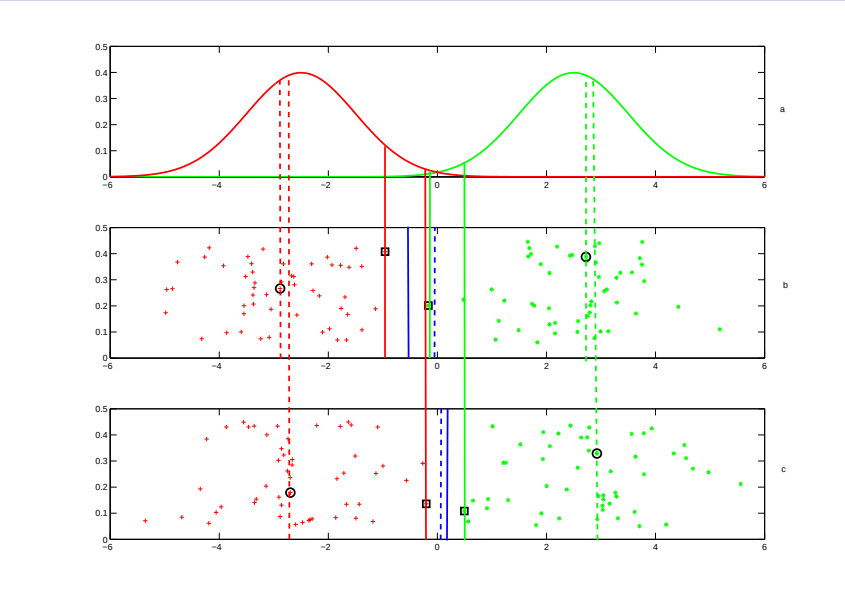
$$\max_{\|w\|=1,b} \quad \{t_{\mathbf{I}}(w,b) + t_{\mathbf{II}}(w,b)\},$$

where

$$t_{\mathbf{I}}(w,b) = \min_{i\in\mathbf{I}}\{y_i(w^T x_i + b)\} \text{ and } t_{\mathbf{II}}(w,b) = \min_{i\in\mathbf{II}}\{y_i(w^T x_i + b)\}.$$

- minimal distance is sensitive to feature noise, and not stable for re-sampling.

minimal distance v.s. quantile distance



- quantile distance is more stable than minimal distance.
- minimizing the pinball loss results in a quantile value,

$$\min_i^{\frac{\tau}{1+\tau}} \{y_i(w^T x_i + b)\} = \arg\min_{t\in\mathbb{R}} \sum_i L_\tau \left(t - y_i(w^T x_i + b)\right),$$

Outline

1. SVM with pinball loss
   - quantile distance
   - pin-SVM
   - properties
2. Solving algorithms
   - primal algorithm
   - sequential minimal optimization
   - solution path
3. Numerical example
4. Conclusion

primal and dual problem

- support vector machine with the pinball loss (pin-SVM[1])

$$\min_{w,b} \ \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m} C_i L_\tau \left(1 - y_i(w^T x_i + b)\right).$$

where

$$L_\tau(u) = \begin{cases} u, & u \geq 0, \\ -\tau u, & u < 0. \end{cases}$$

- the dual problem is

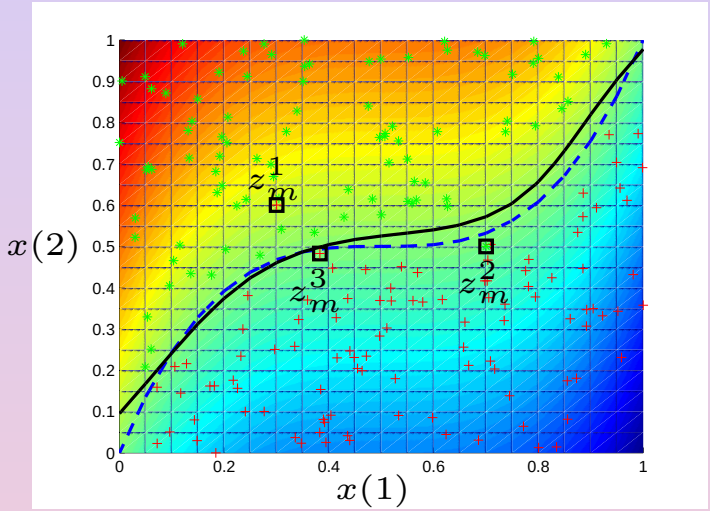
$$\begin{aligned} \max_{\lambda} \quad & -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_i y_i \mathcal{K}_{ij} y_j \lambda_j + \sum_{i=1}^{m} \lambda_i \\ \text{s.t.} \quad & \sum_{i=1}^{m} \lambda_i y_i = 0, \\ & -\tau C_i \leq \lambda_i \leq C_i, i = 1, 2, \ldots, m. \end{aligned}$$

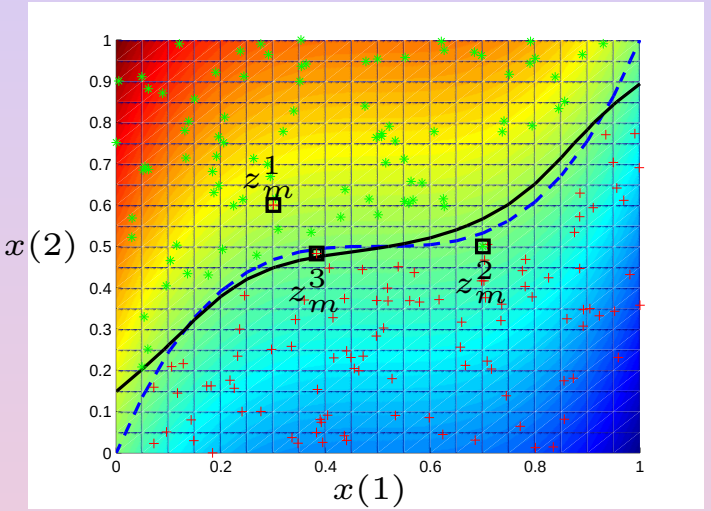
[1] X. Huang, L. Shi, J.A.K. Suykens: Support vector machine classifiers with pinball loss, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(5), 984–997, 2014

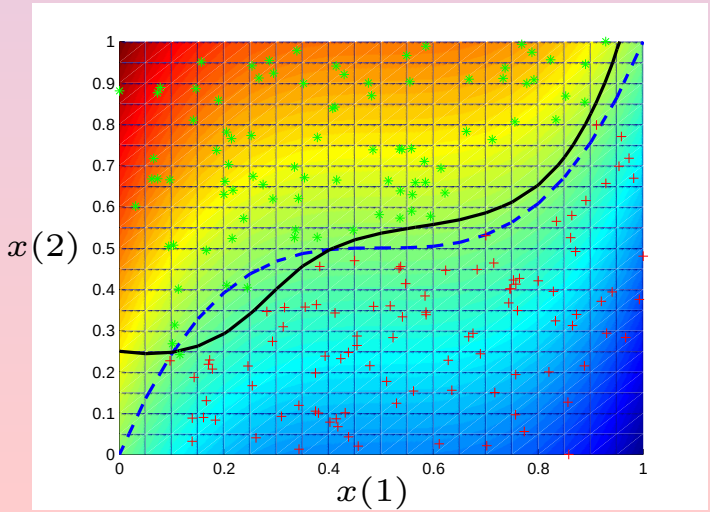
Outline

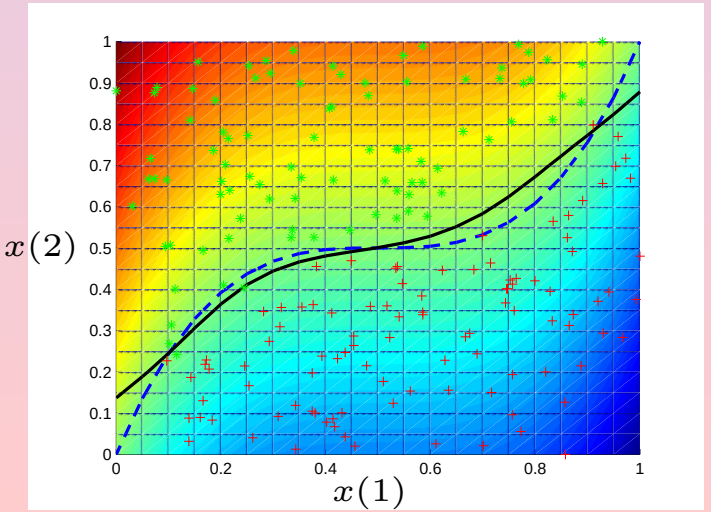
1. SVM with pinball loss
   - quantile distance
   - pin-SVM
   - properties
2. Solving algorithms
   - primal algorithm
   - sequential minimal optimization
   - solution path
3. Numerical example
4. Conclusion

Noise insensitivity









Bayes rule & classification error

Theorem 1

Function $f_{L_\tau,\rho}$, which minimizes the expected L_τ-risk over all measurable functions $f : X \to Y$, is equal to the Bayes classifier, i.e., $f_{L_\tau,\rho}(x) = f_c(x), \forall x \in X$.

Theorem 2

For any probability measure ρ and any measurable function $f : X \to \mathbb{R}$,

$$\mathcal{R}_{L_{\text{mis}},\rho}(\text{sgn}(f)) - \mathcal{R}_{L_{\text{mis}},\rho}(f_c) \leq \mathcal{R}_{L_\tau,\rho}(f) - \mathcal{R}_{L_\tau,\rho}(f_c),$$

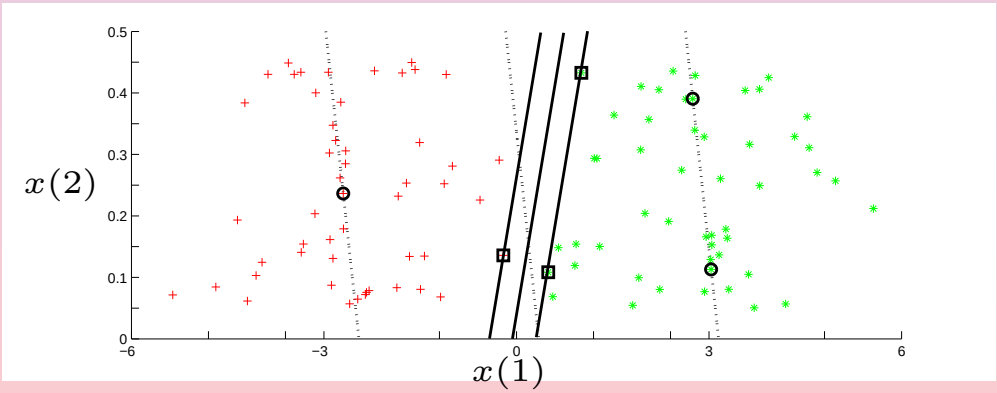
where $\mathcal{R}_{L,\rho}(f) = \int_{X\times Y} L(1 - yf(x))d\rho$.

support vectors

- For obtained λ, b, define the following three sets,

$$\begin{aligned}
\mathcal{L}(\lambda, b) &= \{i : y_i f(x_i) \leq 1, \lambda_i = C_i\}, \\
\mathcal{U}(\lambda, b) &= \{i : y_i f(x_i) \geq 1, \lambda_i = -\tau C_i\}, \\
\mathcal{E}(\lambda, b) &= \{i : y_i f(x_i) = 1, -\tau C_i \leq \lambda_i \leq C_i\}.
\end{aligned}$$

- upper bounded, lower bounded, and free support vectors



Outline

1. SVM with pinball loss
   - quantile distance
   - pin-SVM
   - properties
2. Solving algorithms
   - primal algorithm
   - sequential minimal optimization
   - solution path
3. Numerical example
4. Conclusion

Outline

1. SVM with pinball loss
   - quantile distance
   - pin-SVM
   - properties
2. Solving algorithms
   - primal algorithm
   - sequential minimal optimization
   - solution path
3. Numerical example
4. Conclusion

aLS-SVM

- primal problem:

$$\min_{w,b} \ \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m} C_i L_\tau \left(1 - y_i(w^T x_i + b)\right).$$

- quantile distance can be approached by expectile distance, which can be fast calculated
- support vector machine with the squared pinball loss:

$$\min_{w,b} \ \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m} C_i L_\tau^2 \left(1 - y_i(w^T x_i + b)\right).$$

- a trade-off between LS-SVM and L2-SVM, also called the asymmetric least squares support vector machine (aLS-SVM[2]).
- coordinate descent method for large-scale linear aLS-SVMs.

[2] X. Huang, L. Shi, J.A.K. Suykens: Asymmetric least squares support vector machine, *Computational Statistics and Data Analysis*, 70, 395–405, 2014

Outline

1. SVM with pinball loss
   - quantile distance
   - pin-SVM
   - properties
2. Solving algorithms
   - primal algorithm
   - sequential minimal optimization
   - solution path
3. Numerical example
4. Conclusion

SMO

- dual problem:

$$\begin{aligned} \min_{\lambda} \quad & F(\lambda) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i y_i \mathcal{K}_{ij} y_j \lambda_j - \sum_{i=1}^{m}\lambda_i \\ \text{s.t.} \quad & \sum_{i=1}^{m} y_i\lambda_i = 0, \\ & -\tau C_i \le \lambda_i \le C_i, i = 1, 2, \ldots, m. \end{aligned}$$

- $-\tau C_i \le \lambda_i \le C_i \quad \Leftrightarrow \quad A_i \le y_i\lambda_i \le B_i$, where

$$A_i = \begin{cases} -\tau C_i, & y_i = 1, \\ -C_i, & y_i = -1, \end{cases} \qquad B_i = \begin{cases} C_i, & y_i = 1, \\ \tau C_i, & y_i = -1. \end{cases}$$

- optimality condition: $\max\limits_{i:y_i\lambda_i<B_i} y_i g_i^{\lambda} \le \min\limits_{j:y_j\lambda_j>A_j} y_j g_j^{\lambda}$, where

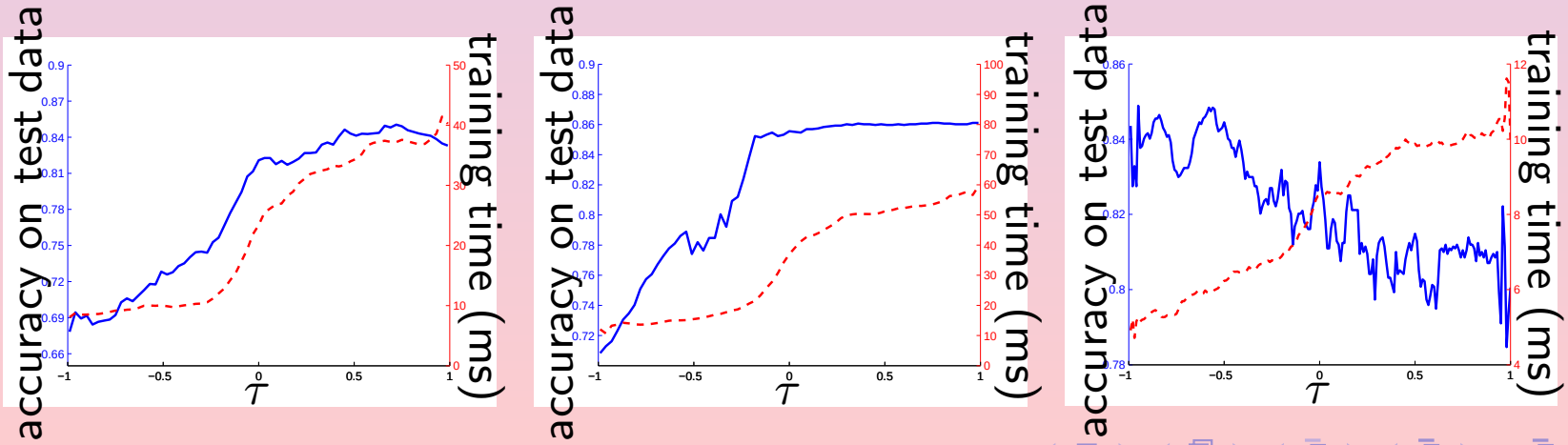
$$g_i^{\lambda} = y_i \sum_{j=1}^{m} y_j\lambda_j\mathcal{K}_{ij} - 1$$

is the gradient of the objective function.

SMO (cont.)

- working set: $i = \arg \max_{i:y_i\lambda_i<B_i} y_i g_i^\lambda$ and $j = \arg \min_{j:y_j\lambda_j>A_j} y_j g_j^\lambda$.
- dual variable update: $\lambda_i = \lambda_i + \xi y_i$ and $\lambda_j = \lambda_j - \xi y_j$
- optimal step length is obtained via solving:

$$\begin{aligned} \min_{\zeta \geq 0} \quad & -\zeta \left(y_i g_i^\lambda - y_j g_j^\lambda \right) + \frac{\zeta^2}{2} \left(\mathcal{K}_{ii} + \mathcal{K}_{jj} - 2\mathcal{K}_{ij} \right) \\ \text{s.t.} \quad & y_i g_i^\lambda + \zeta \leq B_i, \quad y_j g_j^\lambda - \zeta \geq A_j. \end{aligned}$$

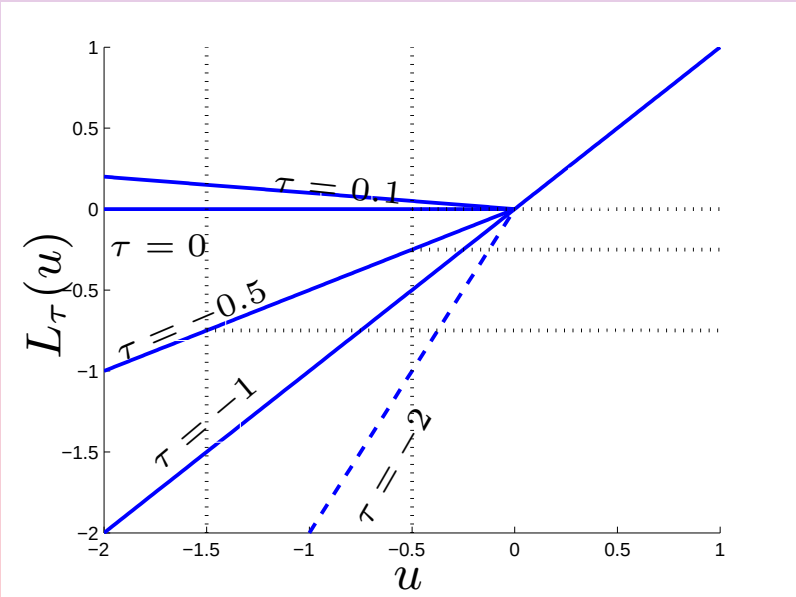


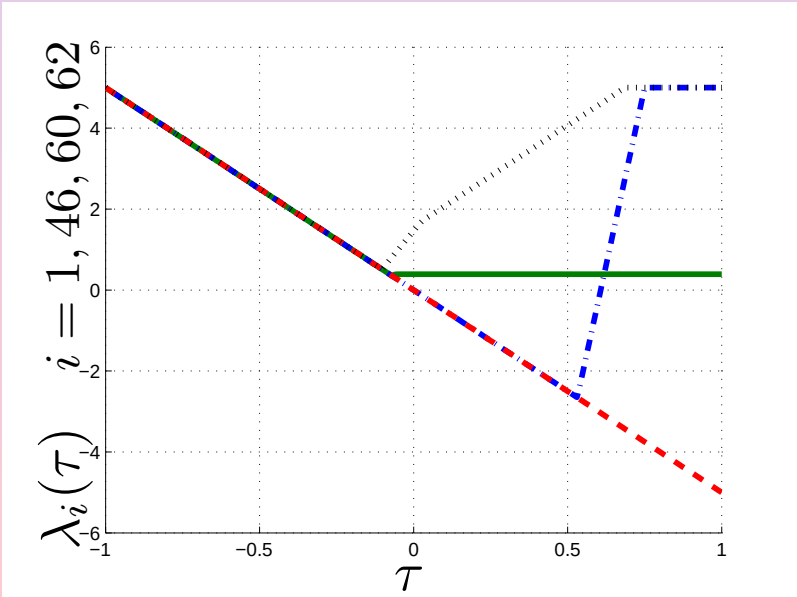
Outline

1. SVM with pinball loss
   - quantile distance
   - pin-SVM
   - properties
2. Solving algorithms
   - primal algorithm
   - sequential minimal optimization
   - solution path
3. Numerical example
4. Conclusion

piecewise linear path

- different τ value is suitable for different problems.
- neither $\tau = 0$ (C-SVM) nor $\tau \geq 0$ is necessary.
- optimal solution is a piecewise linear function w.r.t. τ.





traversal algorithm

- There is a partition $E \subseteq \mathcal{E}(\lambda, b), L \subseteq \mathcal{L}(\lambda, b), U \subseteq \mathcal{U}(\lambda, b)$.
- Optimal solution $\lambda(\tau)$ and $b(\tau)$ are piecewise linear:

$$\begin{aligned} \lambda_i(\tau_0 + \Delta\tau) &= \lambda_i(\tau_0) + a_i \Delta\tau, \\ b(\tau_0 + \Delta\tau) &= b(\tau_0) + a_0 \Delta\tau. \\ y_i f^{\lambda(\tau_0+\Delta\tau), b(\tau_0+\Delta\tau)}(x_i) &= y_i f^{\lambda(\tau_0), b(\tau_0)}(x_i) \\ &\quad + y_i \left(\sum_{j=1}^{m} y_j a_j \mathcal{K}_{ij} + a_0 \right) \Delta\tau. \end{aligned}$$

- suitable $a_i, i = 0, 1, \ldots, m$ should satisfy:
  1. for $i \in E$, the function value keeps unchanged;
  2. for $i \in L$, the dual variable keeps unchanged;
  3. for $i \in U$, the dual variable equals to $-\tau C_i$;
  4. for $i \in E$, the dual variable is in $[-\tau C_i, C_i]$;
  5. for $i \in L$, the function value is not larger than 1;
  6. for $i \in U$, the function value is not smaller than 1.

Outline

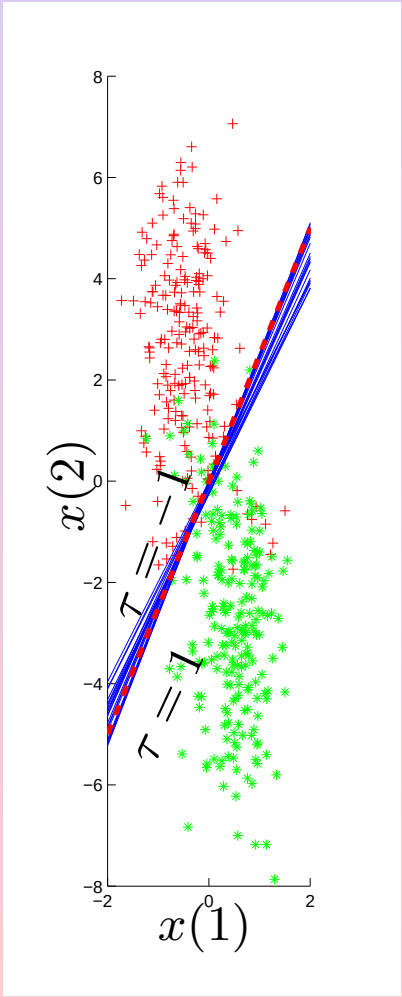
1. SVM with pinball loss
   - quantile distance
   - pin-SVM
   - properties
2. Solving algorithms
   - primal algorithm
   - sequential minimal optimization
   - solution path
3. Numerical example
4. Conclusion

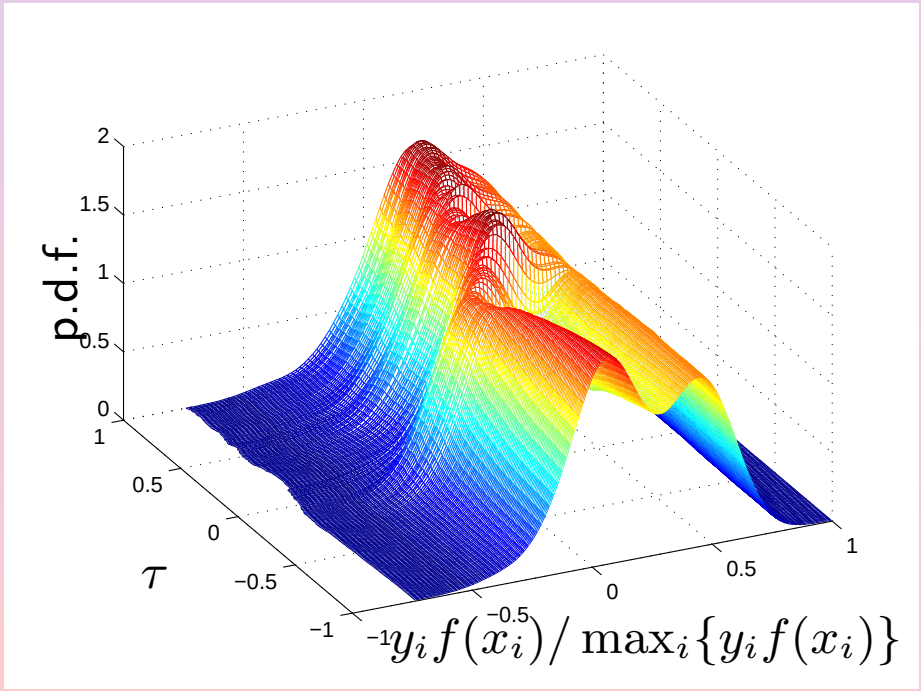
experiments on UCI data set

Table: Test Accuracy, the Suitable τ Values Selected by Cross Validation, and the Computation Time Ratio

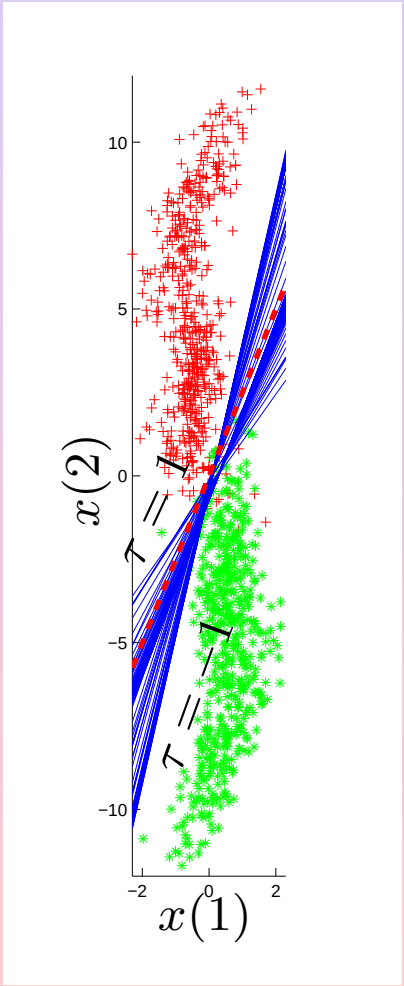
Data	m	C-SVM	pin-SVM	τ	$\frac{T_{\mathrm{tra}}}{T_{\mathrm{SMO}}}$
Spect	80	82.55 ± 2.13	85.53 ± 3.36	-0.44	6.92
Monk1	124	82.18 ± 1.39	84.17 ± 1.24	0.37	2.27
Monk2	169	84.54 ± 0.84	85.12 ± 0.96	0.03	1.98
Monk3	122	92.96 ± 1.02	95.44 ± 1.40	0.28	2.64
Haber.	150	73.32 ± 2.10	73.40 ± 1.78	0.02	3.92
Statlog	150	82.31 ± 2.19	83.00 ± 1.65	-0.11	3.11
Iono.	200	93.22 ± 1.36	93.75 ± 1.75	0.11	3.14
Pima	300	73.20 ± 1.35	74.00 ± 1.97	0.23	3.32
Breast	500	96.78 ± 1.21	97.30 ± 1.19	-0.13	11.8
Trans.	500	74.45 ± 3.48	76.97 ± 2.78	-0.44	6.33
Splice	500	86.51 ± 0.79	86.77 ± 0.53	0.10	6.71
Spamb.	3000	96.70 ± 0.28	96.78 ± 0.04	0.03	20.8
RNA	10000	95.76 ± 0.68	96.03 ± 0.33	-0.05	10.2
IJCNN1	20000	92.62 ± 0.65	92.95 ± 1.13	0.08	28.1

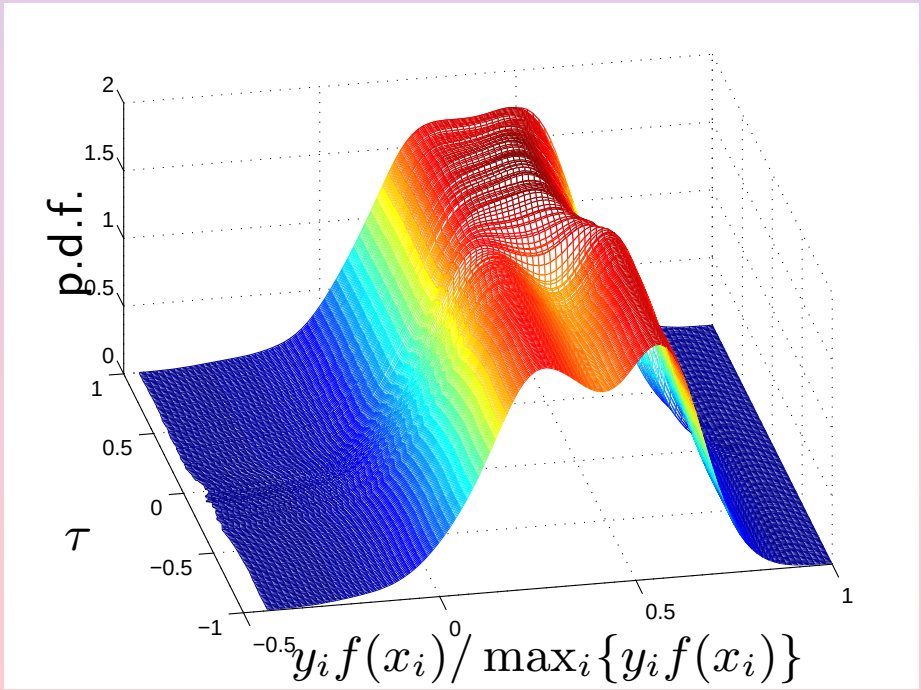
positive τ performs better





negative τ performs better





re-sampling strategy

- data are corrupted by feature noise.
- training data set is small.
- re-sampling and ensemble strategy. (a simple example)
  1. randomly divide the training data into L subsets;
  2. train L sub-classifiers by pin-SVM;
  3. use L predictions made by sub-classifiers for classification.

Table: test accuracy for a simple re-sampling strategy

Data	method	$L = 50$	$L = 20$	$L = 10$
Spamb	C-SVM	86.02 ± 2.80	89.72 ± 2.66	90.07 ± 0.87
	pin-SVM	87.35 ± 2.15	90.74 ± 0.77	91.17 ± 1.33
RNA	C-SVM	94.20 ± 0.63	95.31 ± 0.10	95.43 ± 0.10
	pin-SVM	94.33 ± 0.60	95.43 ± 0.14	95.60 ± 0.12
IJCNN1	C-SVM	86.78 ± 3.98	87.40 ± 3.16	89.33 ± 1.37
	pin-SVM	89.50 ± 1.36	90.01 ± 1.94	90.09 ± 1.64

Outline

1. SVM with pinball loss
   - quantile distance
   - pin-SVM
   - properties
2. Solving algorithms
   - primal algorithm
   - sequential minimal optimization
   - solution path
3. Numerical example
4. Conclusion

conclusion and further study

- statistical analysis of quantile distance for classification and other other problems.
- other convex method for approaching quantile distance.
- learning from the selected τ.
- large-scale linear pin-SVM algorithm.
- re-sampling and ensemble method.

Thanks