Stat 588 – Fall 2007
Data Mining

Lecture 6: Classification and Loss Function

Binary classification notation

- Input: vector $X_i = [X_i[1], \ldots, X_i[p]] \in R^p$
- Output: binary-value $Y_i \in \{-1, 1\}$
- Score: $f(x) \in R$
- Classification rule $h : R^p \rightarrow \{\pm 1\}$:

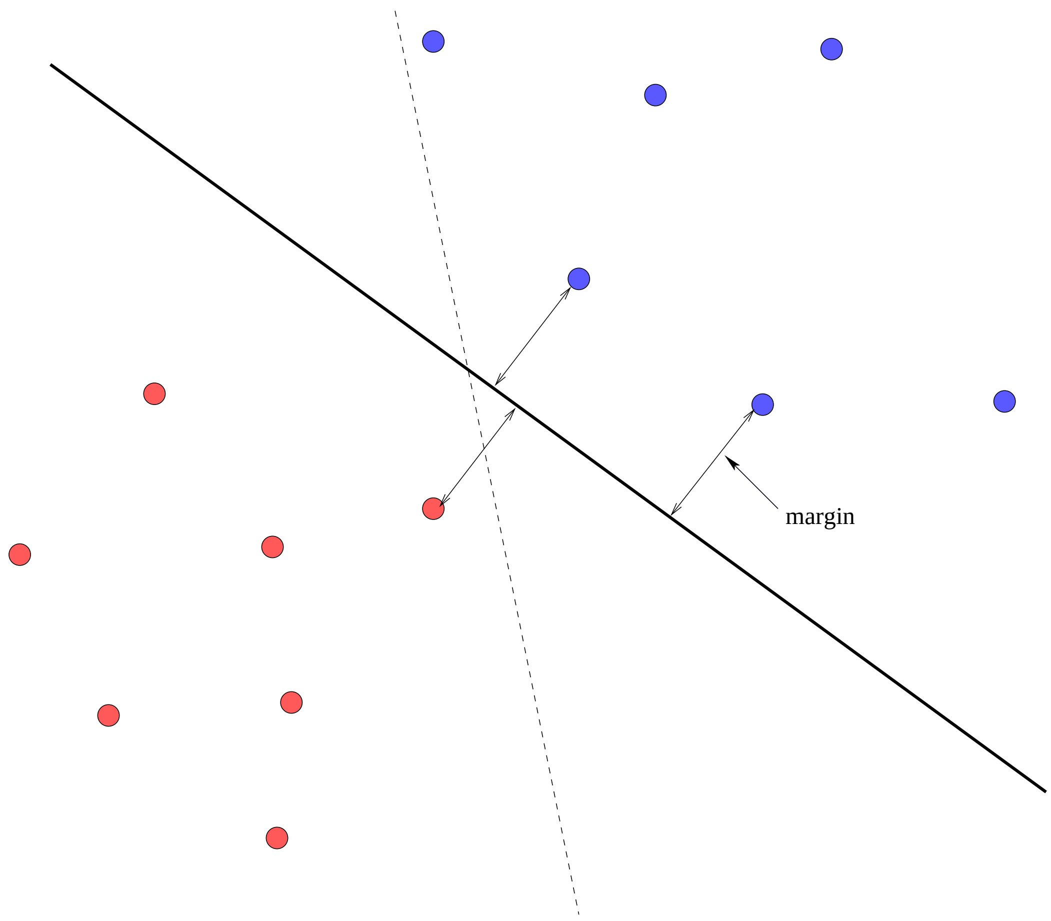
$$h(X) = \begin{cases} 1 & f(X) > 0 \\ -1 & f(X) \leq 0 \end{cases}$$

- Bayes optimal classifier:

$$h_*(X) = \begin{cases} 1 & \text{if } P(Y=1|X) > 0.5 \\ -1 & \text{otherwise} \end{cases}$$

- Methods covered: scoring function $f(x)$ can be trained with: least squares, logistic regression, and perceptron.

Two class linear separator



Optimal separating hyperplane

- Direct maximizing normalized minimum margin:

$$\gamma(w) = \min_i w^T X_i Y_i / \|w\|_2 \sup_i \|X_i\|_2$$

- Convex optimization formulation:

$$\hat{w}_n = \arg\min_w \|w\|_2^2$$

$$\text{subject to } w^T X_i Y_i \geq 1, \quad i = 1, \ldots, n.$$

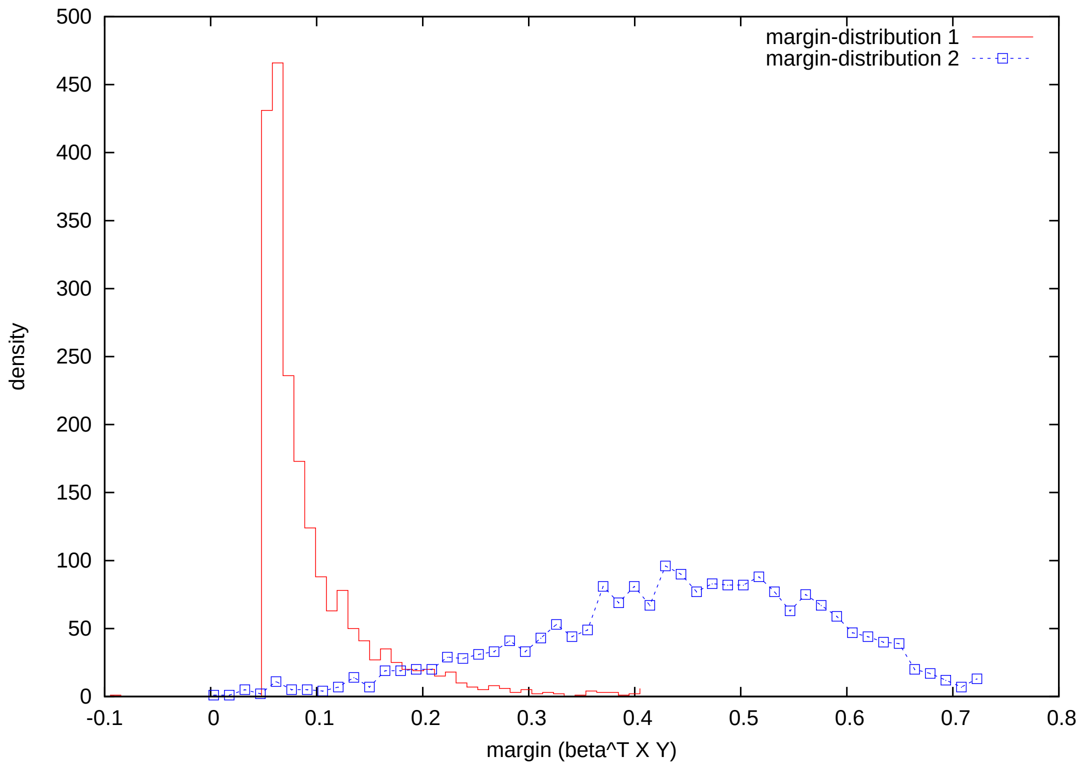
- Support vectors: $\hat{w}^T X_i Y_i = 1$.

- Is minimum margin penalization good criterion?

- Margin bound in the non-separable case: with large probability, the classification error of any data dependent $\hat{w}$ is upper-bounded by

$$\frac{1}{n}\sum_{i=1}^{n} I(\hat{w}^T X_i Y_i \leq \gamma) + O\left(\frac{1}{n}\sqrt{\|\hat{w}\|_2^2 \sum_{i=1}^{n} X_i^2/\gamma^2}\right).$$

Margin distribution



General Linear support vector machine

- Soft-margin penalty function: $C\xi_i$, where

$$(w^T X_i + b)Y_i \geq 1 - \xi_i, \quad \xi_i \geq 0,$$

  and $b \in R$ is called bias.

- Convex optimization formulation:

$$[\hat{w}, \hat{b}] = \arg\min_{w,b} \left[C \sum_{i=1}^{n} \xi_i + \|w\|_2^2 \right]$$

subject to $(w^T X_i + b)Y_i \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, n.$

- Equivalent formulation (by eliminating ξ_i):

$$[\hat{w}, \hat{b}] = \arg\min_{w,b} \left[C \sum_{i=1}^{n} (1 - (w^T X_i + b) Y_i)_+ + \|w\|_2^2 \right].$$

- The SVM loss function:

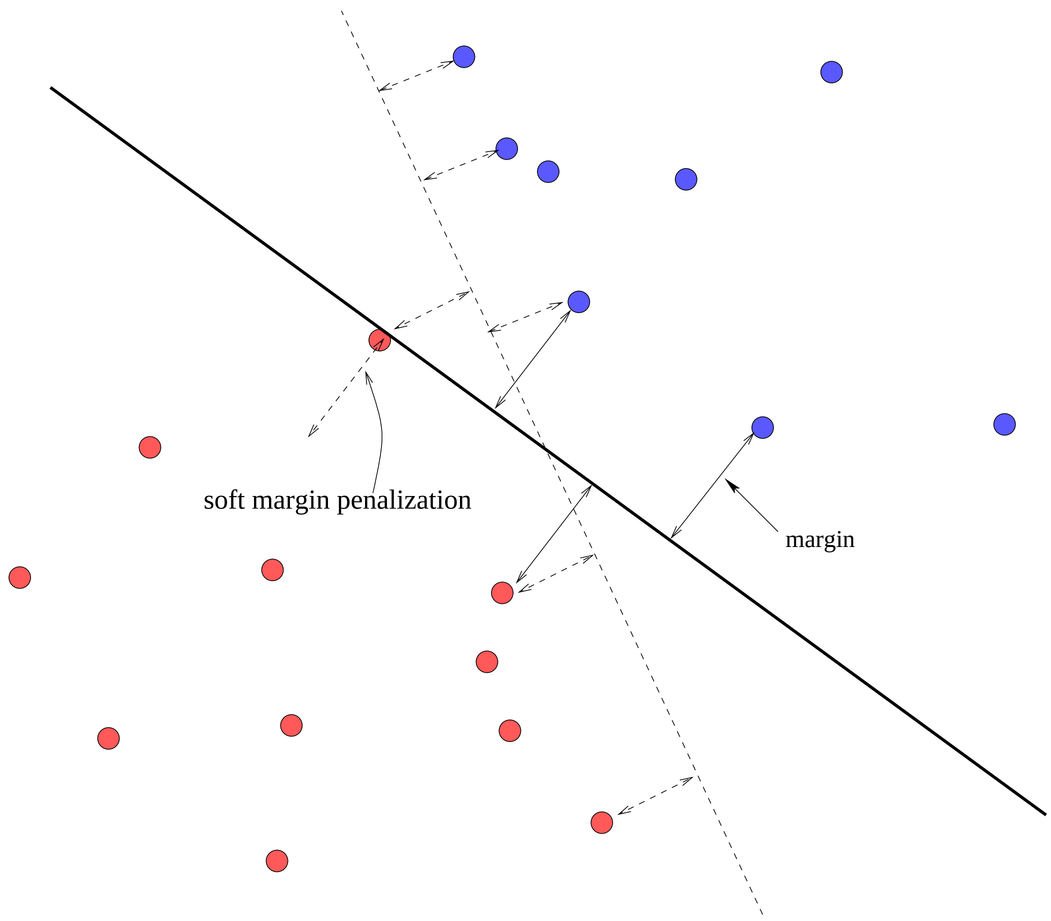
$$L(f, Y) = (1 - fY)_+ = \begin{cases} 1 - fY & \text{if } fY \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

  is also called hinge loss.

- Bias b is not regularized (may remove this parameter, but introduce constant feature 1 into X_i, with corresponding parameter regularized).

- $C \to 0$: the solution goes to optimal separating hyperplane.

SVM: geometric interpretation



General risk minimization framework

- Pick a function class $\mathcal{H}$: e.g. linear function class

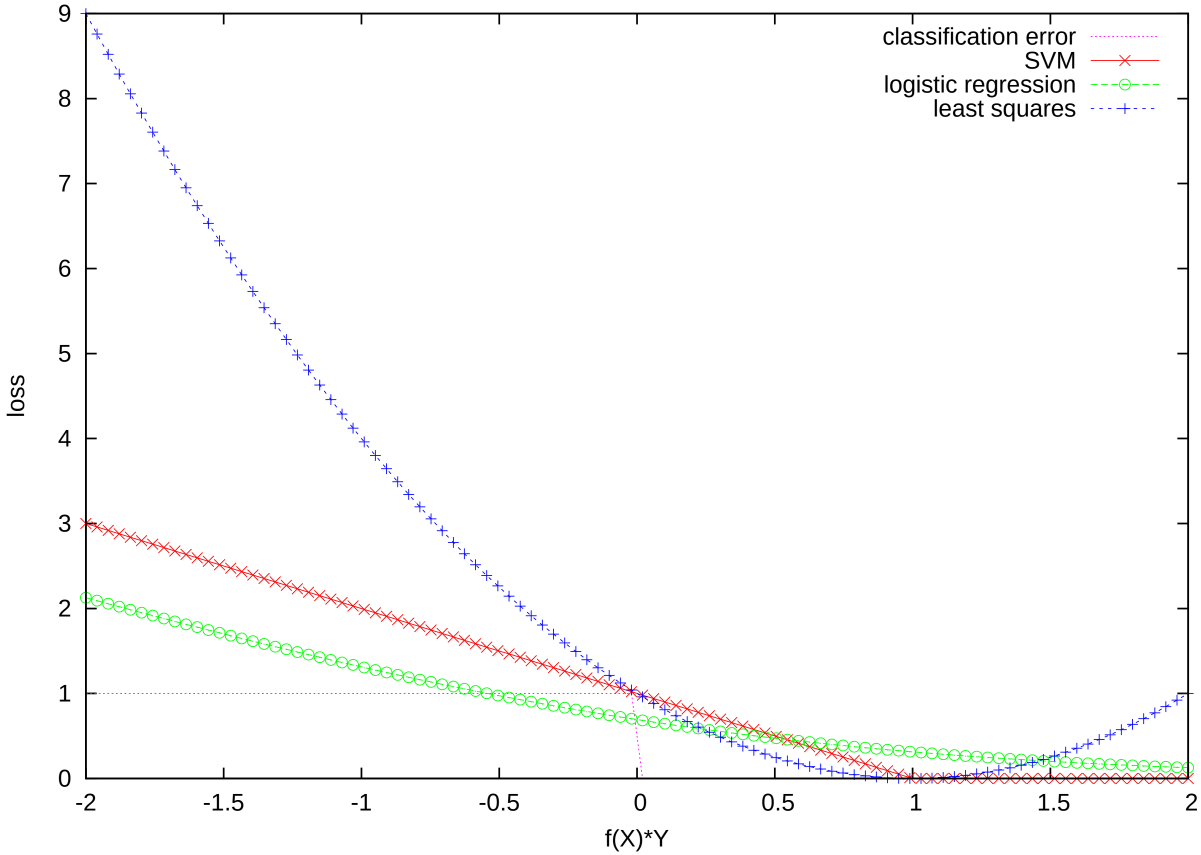
$$\mathcal{H} = \{f(X) = w^T X + b; \|w\|_2^2 \le a\}.$$

- Pick a (convex) loss function $L(f, Y)$, e.g.:

$$L(f, Y) = \phi(fY); \quad \phi(a) = \underbrace{(a-1)^2}_{\text{least squares}}, \underbrace{\ln(1+e^{-a})}_{\text{logistic regression}}, \underbrace{(1-a)_+}_{\text{SVM}}.$$

- find $\hat{f} \in \mathcal{H}$ to minimize empirical error:

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \sum_{i=1}^{n} L(f(X_i), Y_i).$$


classification error
SVM
logistic regression
least squares
loss
f(X)*Y
9
8
7
6
5
4
3
2
1
0
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

Performance Comparison (high dimensional data)

- Text categorization: reuters data, 118 classes, with averaged binary performance over the classes

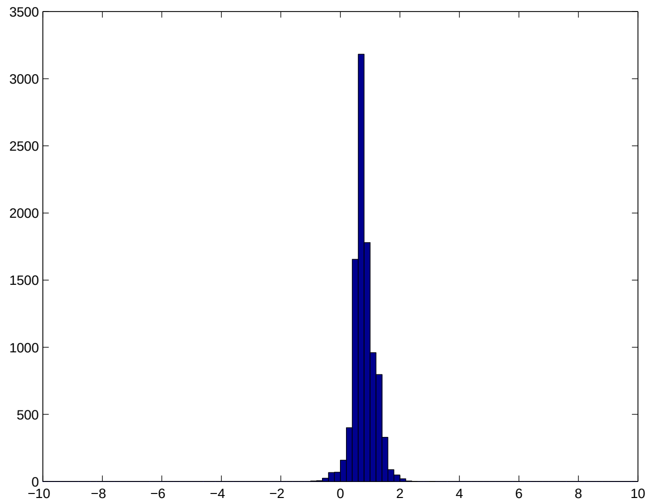
	Naive Bayes	Ridge Reg	Mod LS	Logistic Reg	SVM
precision	77.0	87.1	89.2	88.0	89.2
recall	76.9	84.9	85.3	84.9	84.0
F_1	77.0	86.0	87.2	86.4	86.5

Table 1: Binary classification performance on Reuters (all 118 classes)

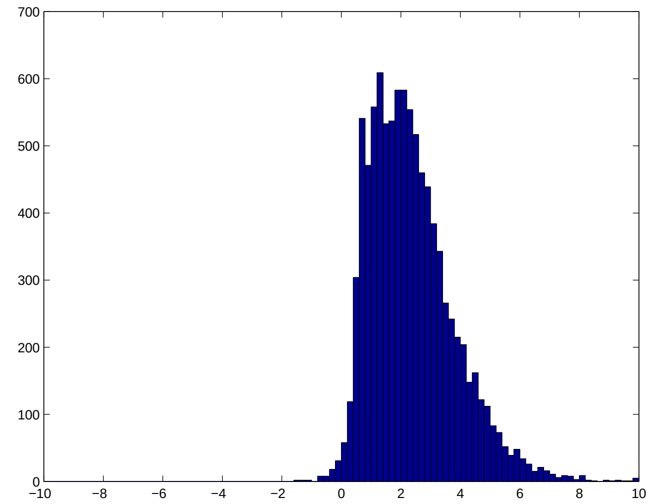
method	IndustrySector	WebKB	GRANT	IBMweb
Naive Bayes-1	84.8 ± 0.5	65.0 ± 0.7	59.4 ± 1.9	77.2 ± 0.4
Naive Bayes-2	91.0 ± 0.6	68.7 ± 1.1	64.2 ± 1.3	79.6 ± 0.7
Lin Reg	93.4 ± 0.5	83.8 ± 0.3	67.0 ± 0.8	85.7 ± 0.5
Mod LS	93.6 ± 0.4	88.7 ± 0.5	70.2 ± 1.2	86.2 ± 0.7
Logistic Reg	92.3 ± 0.9	89.0 ± 0.5	70.6 ± 1.2	86.2 ± 0.6
SVM	93.6 ± 0.5	88.4 ± 0.5	70.0 ± 1.2	86.1 ± 0.4
Mod SVM	93.6 ± 0.4	88.5 ± 0.5	69.8 ± 1.2	85.8 ± 0.7

Table 2: Multi-class classification accuracy

Margin distribution of different loss



least squares method



truncated-least squares method

Figure 1: projected histogram of $\hat{w}^T xy$

Additional loss functions

- Exponential: $L(f(x), y) = \exp(-f(x)y)$.
- Truncated least squares: $L(f(x), y) = \max(1 - f(x)y, 0)^2$.
- ...
- Loss functions determine probability model: probability caliberation.
- Desirable property: binary classification: $f(x) > 0$ equivalent to $P(y = 1|x) > 0.5$.

Convex risk minimization

- Consider $f(X) = w^T X$
- Consider loss function $\phi(f(X), Y)$, and minimizing empirical risk:

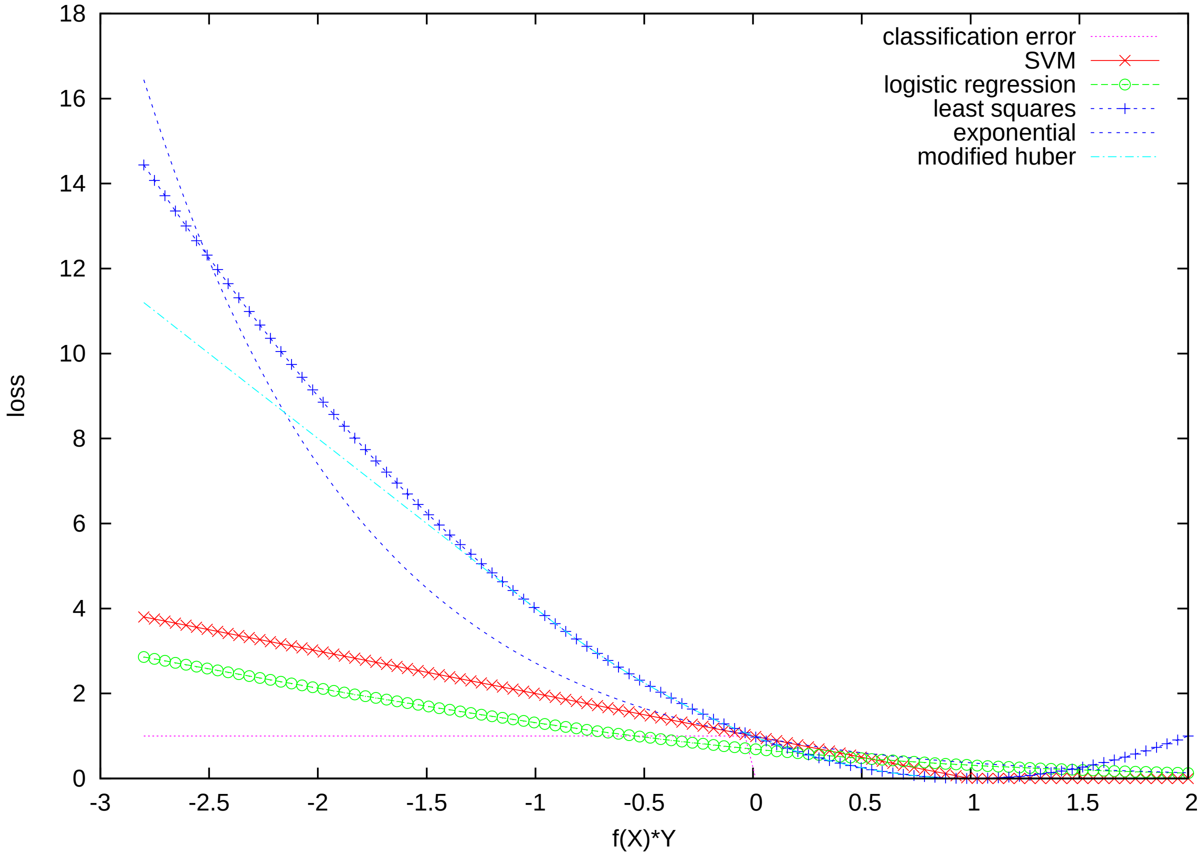
$$\hat{w} = \arg\min_w \sum_{i=1}^n \phi(w^T X_i, Y_i)$$

subject to $g(w) \leq a.$

- Equivalent:

$$\hat{w} = \arg\min_w \sum_{i=1}^n \phi(w^T X_i, Y_i) + \lambda g(w)$$

- Example loss ϕ: least squares, logistic regression, SVM.
- Example regularization $g(w)$: $\|w\|_0$, $\|w\|_1$, $\|w\|_2$.


classification error
SVM
logistic regression
least squares
exponential
modified huber
loss
f(X)*Y
18
16
14
12
10
8
6
4
2
0
-3
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

True risk minimization

- Empirical risk:

$$\hat{w} = \arg\min_w \sum_{i=1}^{n} \phi(w^T X_i, Y_i)$$

subject to $g(w) \leq a$.

- true risk: $f(X) = w^T X$

$$R(f) = \mathbf{E}_{X,Y} \phi(f(X), Y)$$

- $\hat{w}$ approximately minimize true risk:

- $\hat{f}(x) = \hat{w}^T x \approx f_*(x) = \arg\min_f R(f)$.

- true minimizer $f_* = \arg\min_f \mathbf{E}_x \mathbf{E}_{y|x} \phi(f(x), y)$
  - at each point: $f_*(x) = \arg\min_f \mathbf{E}_{y|x} \phi(f(x), y)$.
  - binary classification:

$$f_*(x) = \arg\min_f [p(y = 1|x)\phi(f, 1) + p(y = -1|x)\phi(f, -1)].$$

Classical Examples

- Least Squares:
  - Loss function: $\phi(f, y) = (f - y)^2$
  - True minimizer

$$
\begin{aligned}
f_*(x) =& \arg\min_f [p(y=1|x)(f-1)^2 + p(y=-1|x)(f+1)^2] \\
=& p(y=1|x) - p(y=-1|x).
\end{aligned}
$$

- Logistic Regression:
  - Loss function: $\phi(f, y) = \ln(1 + \exp(-fy))$

– True minimizer

$$
\begin{aligned}
f_*(x) &= \arg\min_f [p(y=1|x)\ln(1+e^{-f}) + p(y=-1|x)\ln(1+e^{f})] \\
&= \ln(p(y=1|x)/p(y=-1|x))
\end{aligned}
$$

Support Vector Machine (SVM)

- Loss function: $\boldsymbol{\phi(f, y) = \max(0, 1 - fy)}$.

- Maximize margin: push positive and negative points apart.

- True minimizer:

$$
\begin{aligned}
f_*(x) =& \arg\min_f [p(y=1|x)\max(0, 1-f) + p(y=-1|x)\max(0, 1+f)] \\
=& 2I(P(y=1|x) > 0.5) - 1.
\end{aligned}
$$

Probability Calibration

- find calibration function c to map score $f(X)$ to $[0, 1]$: $P(Y = 1|f(X)) = c(f(X)))$.

- One dimensional classification problem.

- Goal of calibration: to make score more interpretable.

- Calibration function is often (near) monotonic.

- Calibration generally does not improve classification accuracy.

- Should be performed on hold-out set instead of training set.

Method of calibration

- All conditional density estimation methods.
- Binning (histogram) and kernel methods:

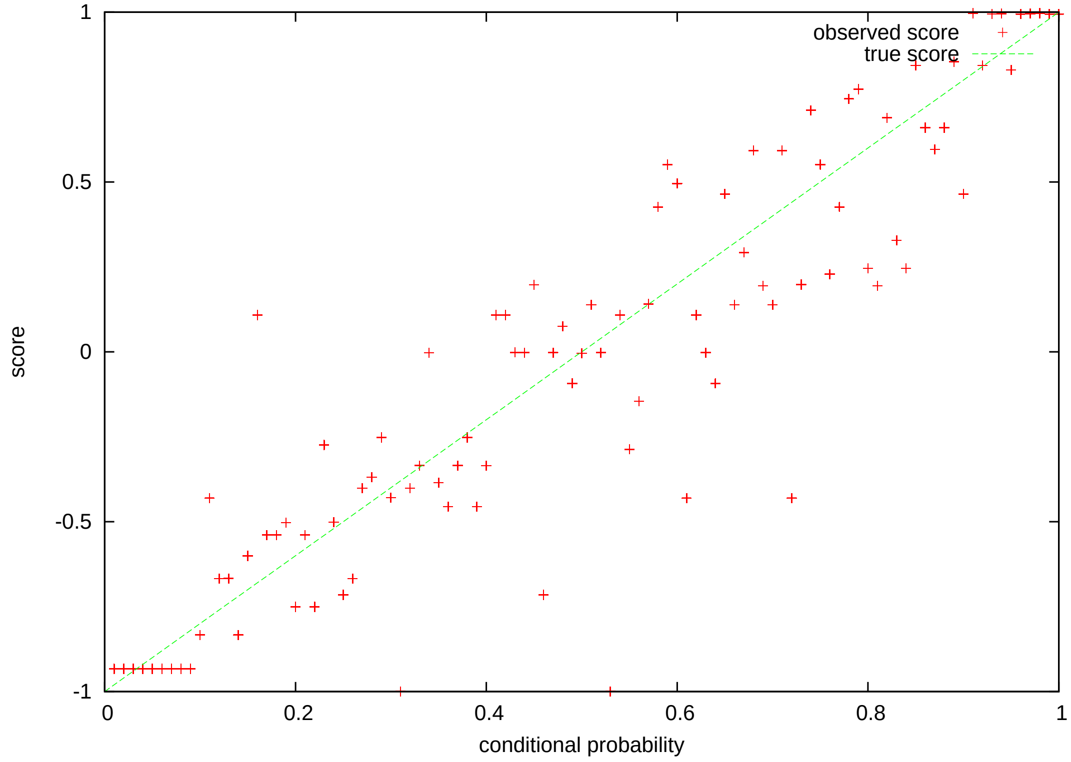
$$c(v) = \frac{\sum_{i:Y_i=1} K(f(X_i), v)}{\sum_i K(f(X_i), v)}.$$

- Logistic regression with basis expansion:

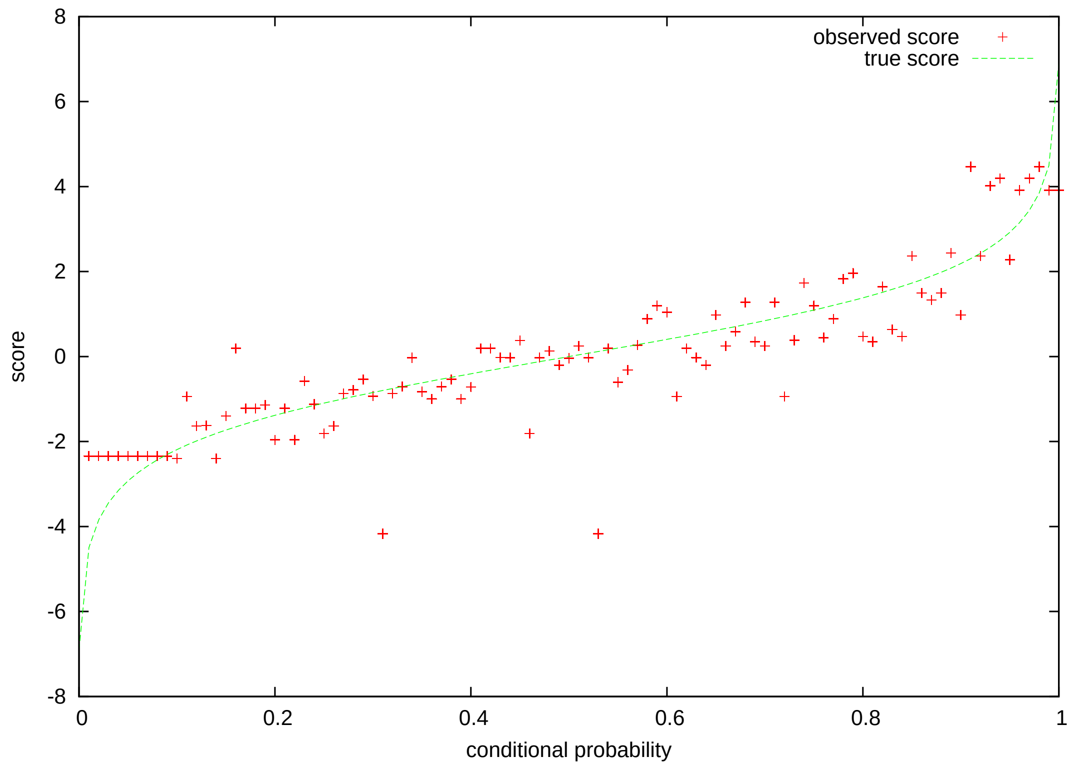
$$c(v) = 1/\exp(a_0 + a_1 v + a_2 h_2(v) + a_3 h_3(v)).$$

use any reasonable basis functions: $h_2(v) = v^2$, $h_2(v) = (v - \xi)_+$, etc.

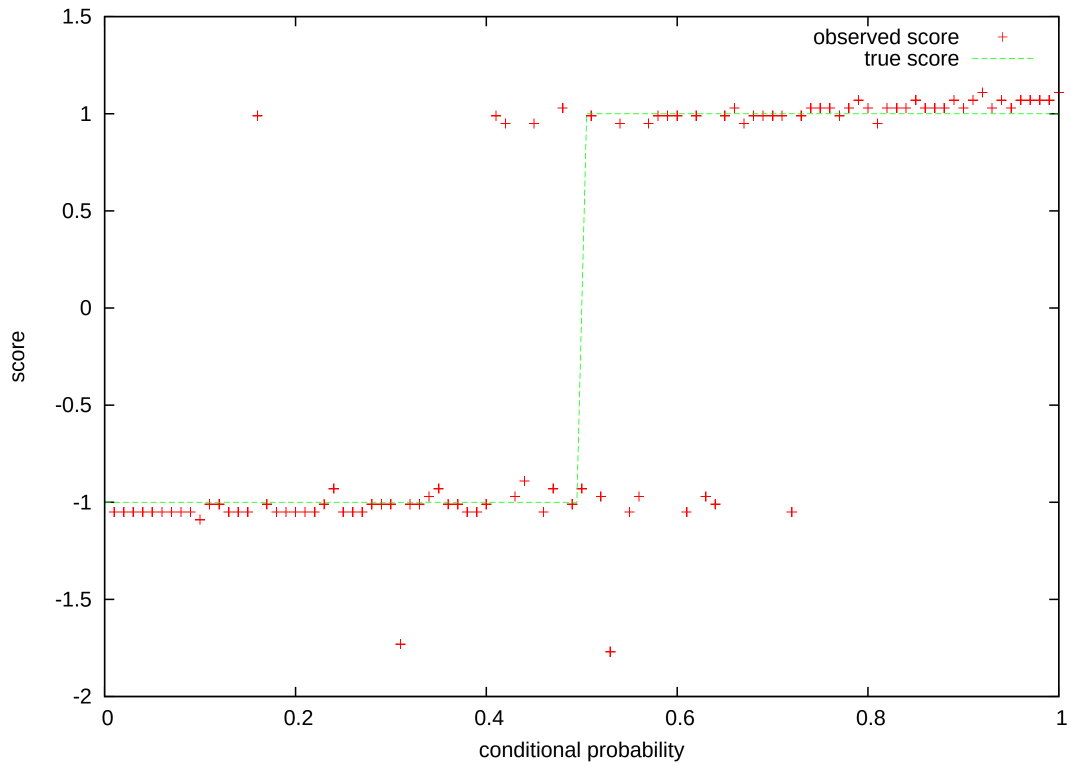
Least Squares



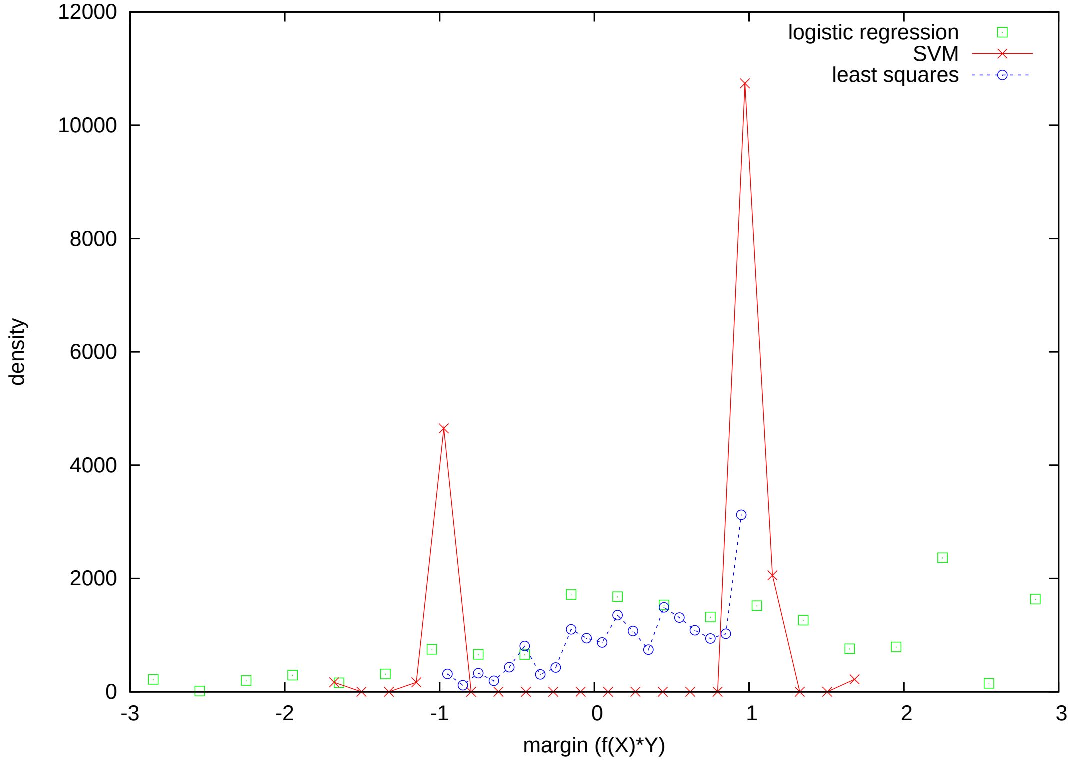
Logistic Regression



SVM


observed score
true score
1.5
1
0.5
0
-0.5
-1
-1.5
-2
score
0
0.2
0.4
0.6
0.8
1
conditional probability

SVM



logistic regression
SVM
least squares
density
margin (f(X)*Y)
12000
10000
8000
6000
4000
2000
0
-3
-2
-1
0
1
2
3

Exponential (Adaboost) Loss

- Loss: $\phi(f, y) = \exp(-fy)$.
  - Optimal minimizer:
    - $\eta = P(y = 1|x)$.
    - $f_\phi^*(\eta) = \frac{1}{2} \ln \frac{\eta}{1-\eta}$.
- Probability model: $f \longrightarrow \bar{\eta} = \frac{1}{1+e^{-2f}}$.

Truncated Least Squares

- Loss: $\phi(f, y) = \max(0, 1 - fy)^2$
  - Optimal minimizer: $f_\phi^*(\eta) = 2\eta - 1$.
- Probability model:

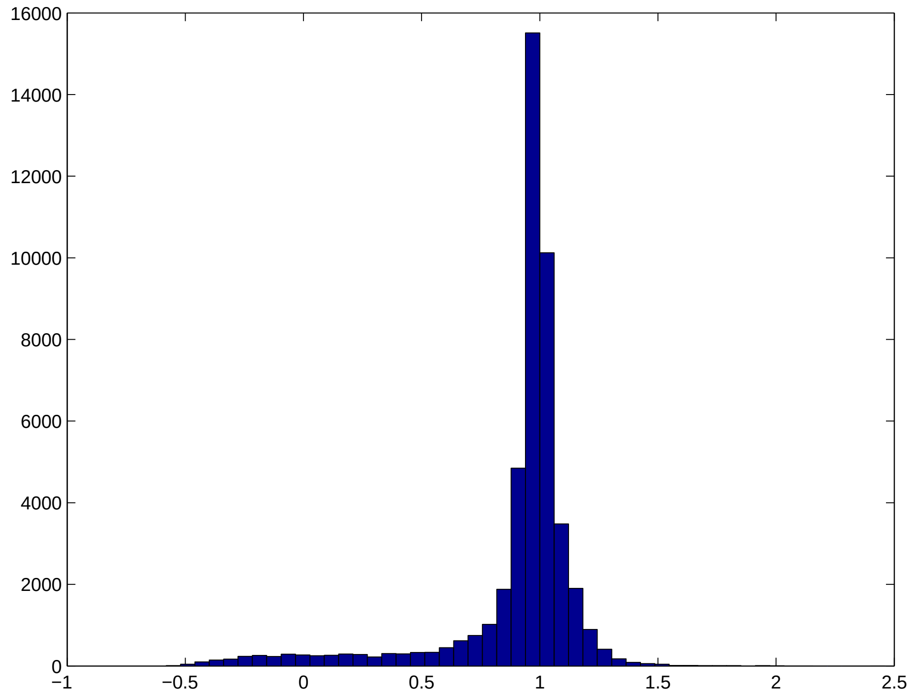
$$f \rightarrow \bar{\eta} = T(f), \qquad T(f) = \min(1, \max(0, (1 + f)/2)).$$

Modified Huber loss

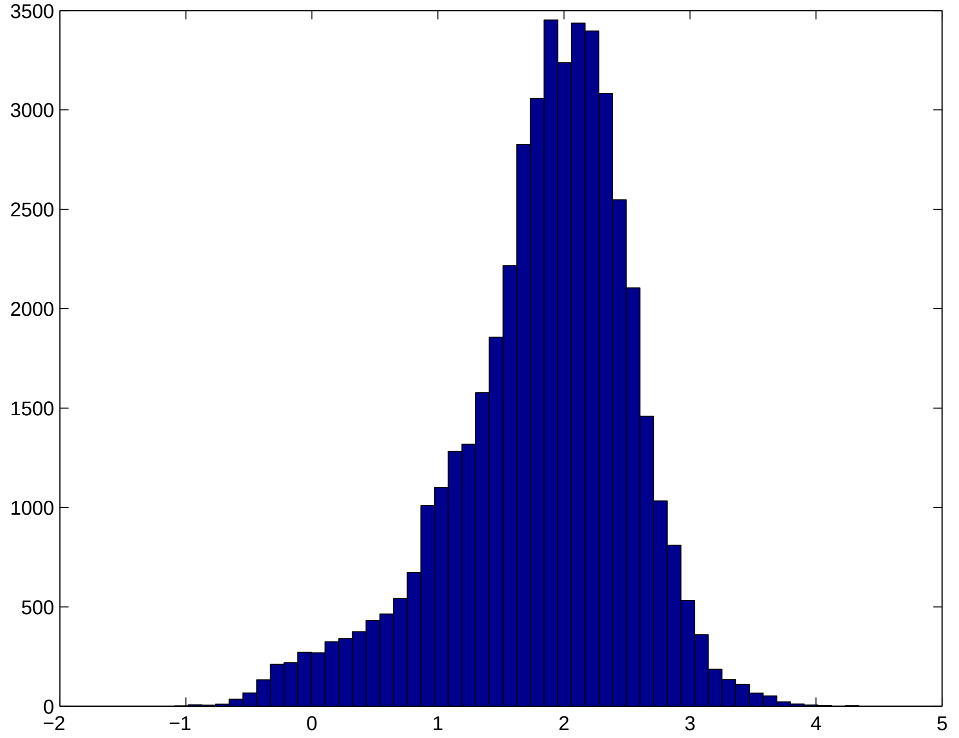
- Loss: $\phi(v) = \begin{cases} -4v & v < -1, \\ (v-1)^2 & v \in [-1, 1], \\ 0 & v > 1. \end{cases}$
  - Optimal minimizer: $f^*_\phi(\eta) = 2\eta - 1$.
- Probability model:

$$f \rightarrow T(f), \qquad T(f) = \min(1, \max(0, (1+f)/2)).$$

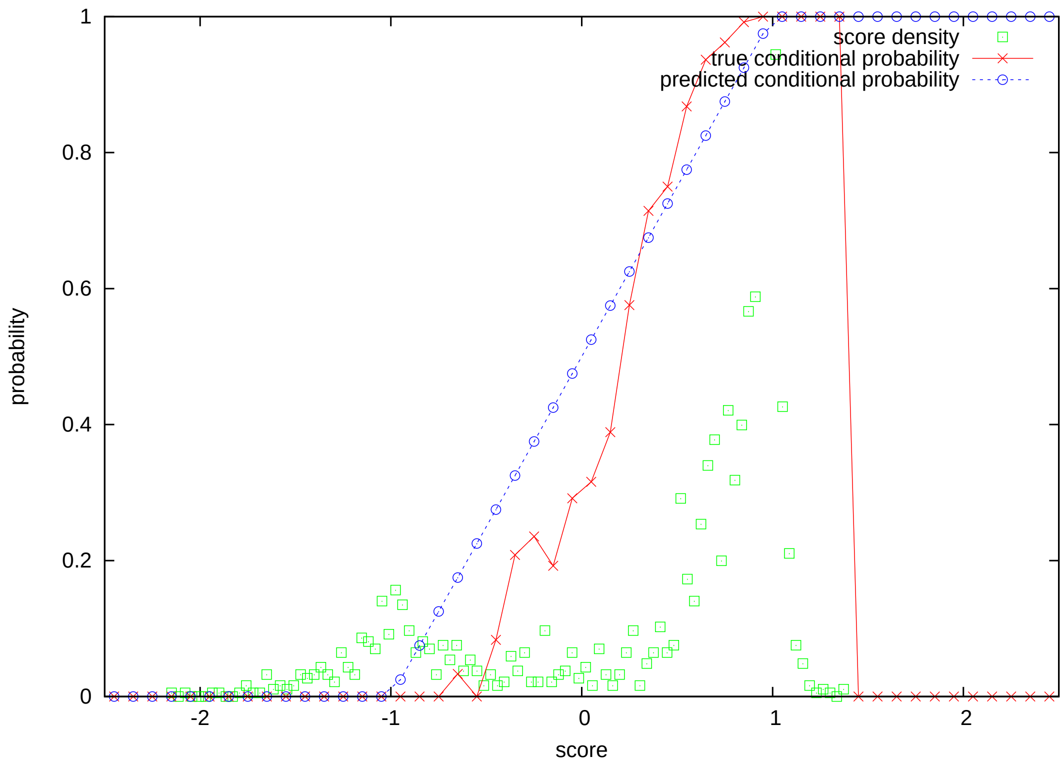
Real data normalized margin-distribution: least squares



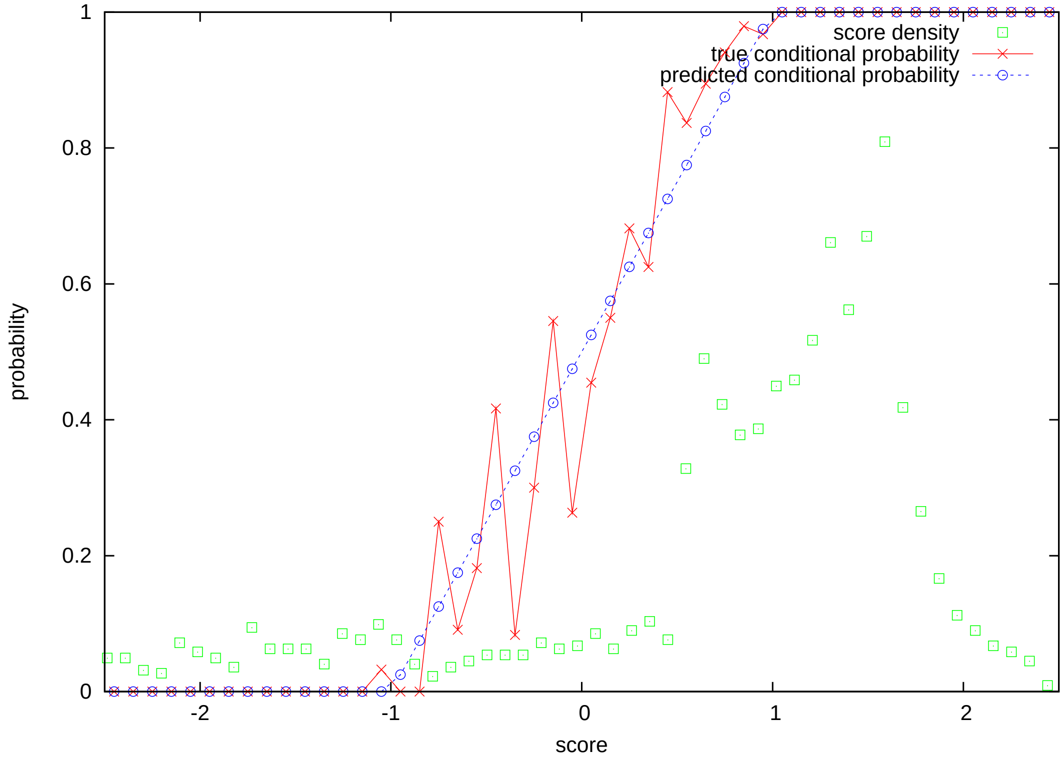
real data normalized margin-distribution: modified Huber



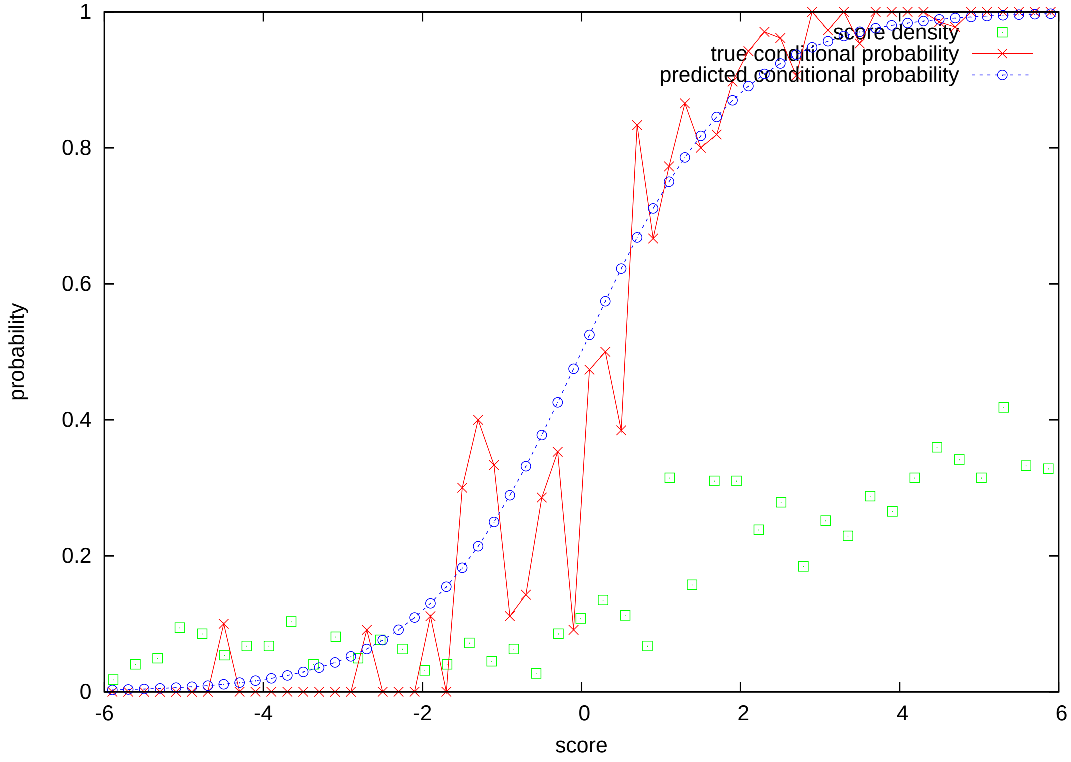
Real data example: least squares



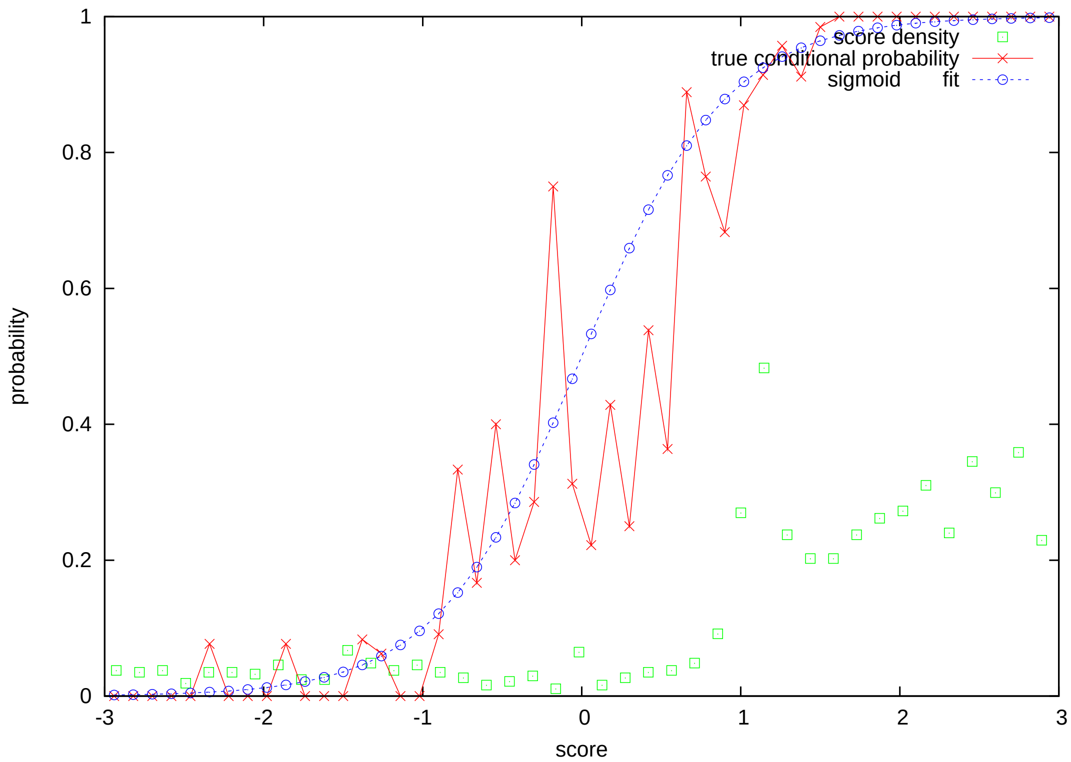
Real data example: truncated least squares



Real data example: logistic regression



Real data example: SVM



References

- Mainly follow

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32:56–85, 2004.

- Also see

P. Bartlett, M. Jordan, and J. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*,