The Convex Upper Bounds of Classification Error Function

陶 卿　2007.11

Qing.tao@mail.ia.ac.cn

中国科学院自动化研究所

Machine Learning and Data Mining 2007

Statistical Machine Learning

- There is a lot of learning algorithms, each method has its own situations for which it works best
- One of the main tasks is to understand the mechanism of each algorithm in a statistical framework.

Contents

- Considering the contrast between the optimal Bayes classifier and classifiers using a classification algorithm
- The closeness is characterized by the convex upper bounds of classification error function
- This is a key step to statistically analyze the learning algorithms.

The Outline

- Some comments on margin-based generalization bound
- The convex upper bounds of classification loss function
- The statistical behavior of different loss functions
- Application in cost-sensitive classification problems

Main References

- T. Zhang. Statistical behaviour and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32:56-85, 2004.
- P. L. Bartlett, et al. Convexity, classification, and risk bounds. *Journal of the American Statistical Association.* 101(473):138-156, 2006.
- I. Steinwart. How to compare *different* loss functions and their risks. *Constructive Approximation.* 26 (2): 225-287 2007.

The Outline

- Some comments on margin-based generalization bound
- The convex upper bounds of classification loss function
- The statistical behavior of different loss functions
- Application in cost-sensitive classification problems

VC Theory and PAC Bounds

- Landmark paper by Blumer, 1989;
  - Greatly influence the field of machine learning
  - VC theory and PAC bounds have been used to analyze the performance of learning systems as diverse as decision trees, neural networks, and others.

PAC Bound: margin

- The first paper about margin results: J.Shawe-Taylor and P. L. Bartlett, 1998.

$$err_D(h) \le \varepsilon(l, F, \delta, \gamma) \le \frac{2}{l}\left(\frac{64R^2}{\gamma^2}\log\frac{el\gamma}{8R^2}\log\frac{32l}{\gamma^2} + \log\frac{4}{\delta}\right)$$

provided $\frac{64R}{\gamma^2} < l \qquad l > 2/\varepsilon$

- Data dependent and dimension free.

Margin-based Bound

- During 1992-2004, the most influential explanation is the so-called “margin” analysis. This concept has been used to explain both SVM and boosting.

Three Periods of Inference Science

- 1970 – 1990 Development of Basics of Statistical Learning Theory (the VC theory)
- ***1992 – 2004 Development of Large Margin Technology (SVMs)***
- 2005 – ....Development of Non-Inductive Methods of Inferences.

Comment

- J. H. Friedman, T. Hastie and R. Tibshirani. Additive Logistic Regression: a Statistical View of Boosting. *The Annals of Statistics*. 2000, 28(2): 337-407.

- The bounds and the theory associated with the AdaBoost algorithms are interesting, but tend to be too loose to be of practical importance.

Comment

- I. Steinwart. 2002

- Although the existing bounds are usually too large for real-word sample sizes, it is claimed that at least for large sample sizes these bounds can justify SVM approach. In the presence of noise, many of the known bounds can not predict well neither for small nor for large sample sizes.

Comment

- T. Zhang 2004

- In statistical estimation procedure, one typically encounter two types of errors: approximation error and variance estimation error. The margin idea mixes this two aspects together. It is not clear that which aspect is the main contribution to the success of maximum margin methods. Moreover, the impact of different loss functions can not be characterized.

Comment

- Trevor Hastie and Ji Zhu 2006
- What is special with the SVM is not the regularization term, but is rather the loss function, that is, the hinge loss.
- The hinge loss and other loss functions of many statistical tools are all Bayes consistent. This fact justifies that margin maximization is not the key to the success of the SVM.

The Outline

- Some comments on margin-based generalization bound
- The convex upper bounds of classification loss function
- The statistical behavior of different loss functions
- Application in cost-sensitive classification problems

Definition of Classification

- Assumption： (x_i, y_i) i.i.d.
- Hypothesis space： H
- Loss function： $$c(y, f(x)) = \begin{cases} 0 & if \quad f(x) = y \\ 1 & if \quad f(x) \neq y \end{cases}$$
- Objective function： $$R(f) = \int c(y, f(x)) P(x, y) dx$$

Bayesian Classifier

- H: all measurable functions

- The optimal classifier is

$$f^*(x) = \begin{cases} 1 & if \quad \eta(x) > \frac{1}{2} \\ -1 & if \quad \eta(x) < \frac{1}{2} \end{cases}$$

Classification in Machine Learning

- Only finite samples are available
- Obviously, minimizing expected risk analytically is impossible
- Instead, we can minimize the empirical risk.

Direct Optimization

- The empirical risk

$$\min \frac{1}{n} \sum_{i=1}^{n} c\left(y_i, f\left(x_i\right)\right)$$

- In fact, Ben-David et al. [2003] show that even approximately minimizing the empirical risk is NP-hard, not only for linear function classes but also for spheres and other simple geometrical objects.

Intuitive Convex Upper Bound

- $\phi : \mathrm{R} \to [0, \infty)$ is a convex function
- There exists a $\gamma > 0$, such that
  $$\gamma\phi(yf(x)) \geq c(y, f(x)) \quad \text{for all } x \in R$$
- They have the same optimal classification solutions.

Computational Heuristics

- Many of the most prominent methods studied in machine learning make significant use of convexity
- Generally, the margin-based surrogate function is to be minimized to obtain a classifier

Margin-based Loss functions

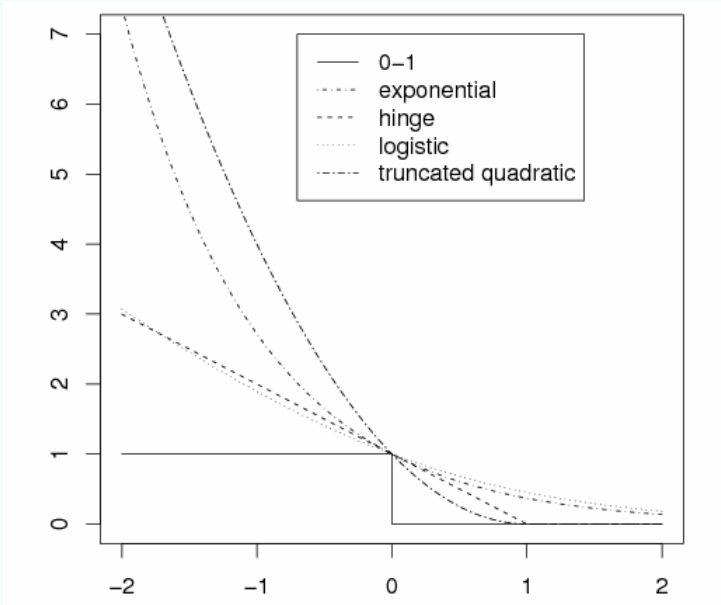
- Margin: $v = yf(x)$
- Modified Least Squares: $\phi(v) = \max\{1-v, 0\}^2$
- SVM: $\phi(v) = \max\{1-v, 0\}$
- Exponential: $\phi(v) = \exp(-v)$
- Logistic Regression: $\phi(v) = \ln(1+\exp(-v))$

Empirical Consistency

- Although SVM and AdaBoost don't directly optimize the classification error, great empirical success has been achieved.
- So, such surrogate loss functions must be “reasonably related” to the original loss function since otherwise this approach cannot work well.

Illustration

- The convex upper bounds of classification error function or “surrogate” functions.



Conditions on the Surrogate

- Conditions such as convexity, continuity, and differentiability of ϕ are easy to verify and have natural relationships to optimization procedures, it is not immediately obvious how to relate such conditions to their statistical consistency between the 0-1 loss and its surrogate loss .

- *Thus, what condition on ϕ should be further considered?*

Understanding the Convex Upper Bounds

- It is very important that we should understand these strategies not only from a computational point of view but also in terms of their statistical properties.

Questions

- *What are their Bayesian solutions ?*

- *Under what conditions, does the convergence of convex bound risk imply the convergence of original risk?*

Some Notations

- ϕ-risk in binary classification

$$R_\phi(f) = EQ(\eta(x), f(x)) \qquad Q(\eta, f) = \eta\phi(f) + (1-\eta)\phi(-f)$$

- Optimal ϕ-classifier

$$f_\phi^*(\eta) = \arg\min_{f \in R^*} \{\eta\phi(f) + (1-\eta)\phi(-f)\}$$

$$H(\eta) = \inf_{f \in R} \{\eta\phi(f) + (1-\eta)\phi(-f)\} = Q(\eta, f_\phi^*(\eta))$$

- Excess ϕ-risk

$$\Delta Q(\eta, f) = Q(\eta, f) - H(\eta)$$

Minimizing the Risk

- The minimal ϕ-risk can be achieved by pointwisely minimizing

- $$Q(\eta, f) = \eta\phi(f) + (1-\eta)\phi(-f)$$

Partial Answer

- Modified Least Squares: $f_\phi^*(\eta) = 2\eta - 1$, $H(\eta) = 4\eta(1-\eta)$
- SVM: $f_\phi^*(\eta) = sign(2\eta - 1)$ $H(\eta) = 1 - |2\eta - 1|$
- Exponential: $f_\phi^*(\eta) = \frac{1}{2}\ln\frac{\eta}{1-\eta}$ $H(\eta) = 2\sqrt{\eta(\eta-1)}$
- Logistic Regression: $f_\phi^*(\eta) = \ln\frac{\eta}{1-\eta}$ $H(\eta) = -\eta\ln\eta - (1-\eta)\ln(1-\eta)$
- *Bayes consistent or Fisher consistent for classification problems*

Fisher Consistency

- In the traditional parameter estimation situation, Fisher consistency means that the estimation procedure in the population space will produce the target of the estimation.

- The Fisher consistency of the margin-based loss functions is closely related to the consistency and rate of convergence (to the Bayes optimal risk) results of the corresponding classifiers (Y. Lin 2004).

Classification-calibrated

- A loss function is classification-calibrated if,

$$H^{-}(\eta) > H(\eta) \qquad \text{for any} \quad \eta \neq \frac{1}{2}$$

$$H^{-}(\eta) = \inf_{v(2\eta-1)\leq 0} \eta\phi(v) + (1-\eta)\phi(-v)$$

- This is a minimal condition that can be viewed as a pointwise form of Fisher consistency for classification (Bartlett, et al 2006).

Partial Answer

The following two conditions are equivalent (Bartlett, et al 2006):

- *The convergence of ϕ–excess risk implies the convergence of original excess risk*
- *The surrogate function ϕ is classification-calibrated.*

Importance

- Transfer assessments of statistical error in terms of "excess ϕ-risk" $R_\phi(f) - R_\phi^*$ into assessments of error in terms of "excess risk" $R(f) - R^*$

- Under the condition of classification-calibrated, the surrogate loss functions are "reasonably related" to the original loss function.

Example: classification-calibrated

- Hinge loss function is classification-calibrated.

Convexity

- Let the loss function ϕ be convex. Then

  it is classification-calibrated if and only if it is differentiable at 0 and $\phi'(0) < 0$

- Exponential and logistic loss are classification-calibrated.

Upper Bound

- *If* $\phi : \mathrm{R} \to [0, \infty)$ *is classification-calibrated, then there exists a* $\gamma > 0$ *, such that*

$$\gamma\phi(v) \geq I(v \leq 0) \qquad \text{for all } v \in R$$

Quantitative Relationship

- Bartlett et al. simplify and extend Zhang's results, developing a general methodology for finding quantitative relationships

$$\psi\left(R(f)-R^{*}\right)\leq R_{\phi}(f)-R_{\phi}^{*}$$

- where ψ is the Fenchel-Legendre biconjugate of

$$H^{-}\left(\frac{1+\theta}{2}\right)-H\left(\frac{1+\theta}{2}\right)$$

Classifiers and Density Function

$$f_\phi^*(\eta) = 2\eta - 1 \qquad f_\phi^*(\eta) = \frac{1}{2}\ln\frac{\eta}{1-\eta}$$

- The density function can be obtained. This means solving classification is equivalent to estimating density.

Bregman Divergence

- For a convex function, the Bregman divergence is defined as

$$d_\phi(x_1, x_2) = \phi(x_2) - \phi(x_1) - \phi'(x_1)(x_2 - x_1)$$

- If ϕ and f_ϕ is differentiable, then $H(\eta)$ is also differentiable
- If ϕ is convex, then $H(\eta)$ is concave and the Bregman divergence can be uniquely defined.

Classification and Density Estimation

- Assume that $p = f_\phi(\overline{\eta})$, then

$$\Delta Q(\eta, p) = d_H(\overline{\eta}, \eta)$$

- Excess ϕ -risk

$$E_x \Delta Q(\eta(x), f(x)) = E_x d_H\left(f_\phi^{-1}(f(x)), \eta(x)\right)$$

- ***Intuitively, by minimizing the excess-risk, we are effectively minimizing the expected Bregman divergence.***

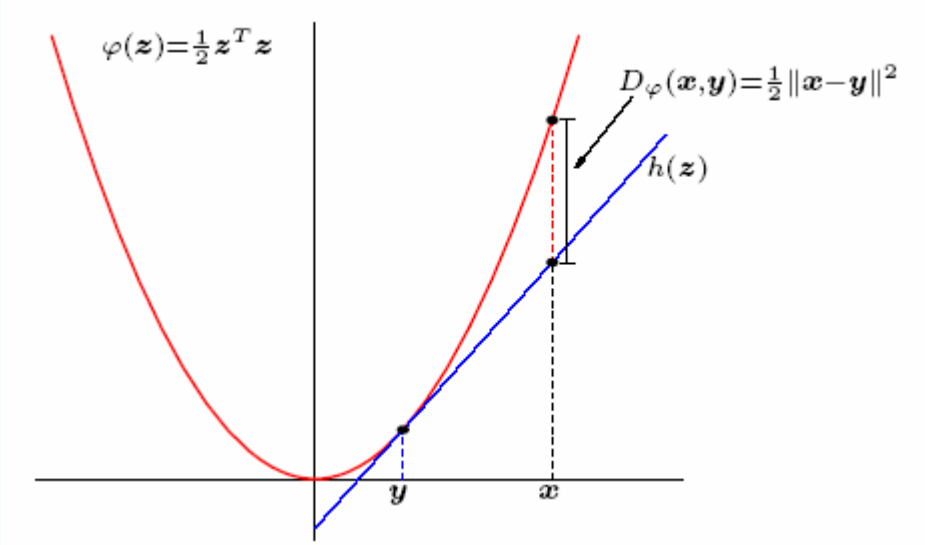
Bregman Divergence

- In mathematics, Bregman divergence is similar to a metric, but does not satisfy the triangle inequality nor symmetry

- Bregman divergence is named after L. M. Bregman, who introduced the concept in 1967. More recently researchers in geometric algorithms have shown that many important algorithms can be generalized from Euclidean metrics to distances defined by Bregman divergence

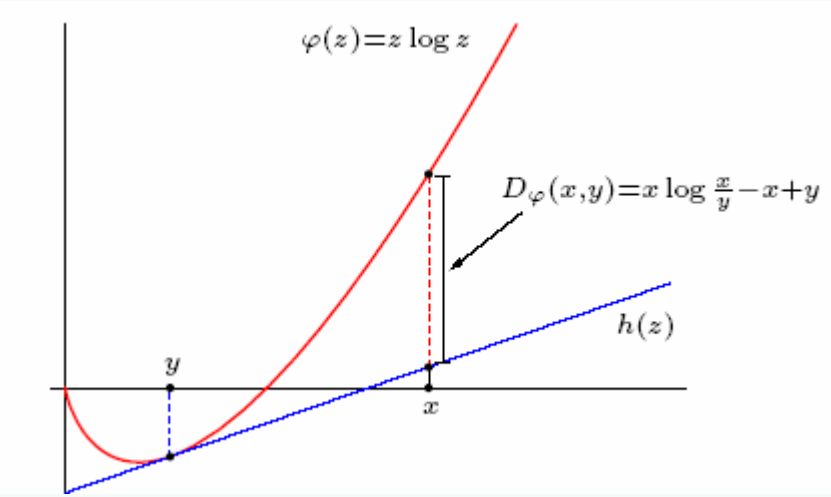
Example: Bregman divergence

- Squared Euclidean distance is a Bregman divergence



Example: Bregman divergence

- Relative Entropy (also called KL-divergence) is a Bregman divergence



The Outline

- Some comments on margin-based generalization bound
- The convex upper bounds of classification loss function
- The statistical behavior of different loss functions
- Application in cost-sensitive classification problems

Statistical Analysis

- *If* $R_\phi(f) - R_\phi^*$ *is small, what kind of statistical behavior it implies?*

- *Mainly by considering*

$$\Delta Q(\eta, p) = \eta\phi(p) + (1-\eta)\phi(-p) - H(\eta)$$

Exponential Loss

- $f_\phi^*(\eta) = \ln \frac{\eta}{1-\eta}$ $\qquad \overline{\eta} = f_\phi^{-1}(p) = \frac{1}{1+e^{-2p}}$
- $\Delta Q(\eta, p) \geq |\eta - \overline{\eta}|\left(e^{|p|} - 1\right) - 2\sqrt{|\eta - \overline{\eta}|}$

  $\Delta Q(\eta, p) \leq |\eta - \overline{\eta}| e^{|p|} + 2\sqrt{|\eta - \overline{\eta}|}$
- $\frac{1}{1+e^{-2f(x)}}$ is regard as an approximation to the true density function $\eta(x)$

Disadvantage

- Using the exponential loss, we compute a predictor such that $|f(x)|$ is large when $\eta(x) = 0,1,$ and $|f(x)|$ is small else where.
- In the limit of zero error, $|f(x)|$ has to achieve $\pm\infty$ if $\eta(x) = 0,1$
- Such a predictor is clearly not well-behaved.

Logistic Loss

- $f_\phi^*(\eta) = \ln \frac{\eta}{1-\eta}$ $\qquad \overline{\eta} = f_\phi^{-1}(p) = \frac{1}{1+e^{-p}}$

- $$\Delta Q(\eta, p) \geq 2(\eta - \overline{\eta})^2$$

- $$\Delta Q(\eta, p) \leq 2|\eta - \overline{\eta}| \ln\left(1 + e^{|p|}\right) + 2\sqrt{k|\eta - \overline{\eta}|}$$

- $\frac{1}{1+e^{-f(x)}}$ is regard as an approximation to the true density function $\eta(x)$

Disadvantage

- In the limit of zero error, there exists the same problem as that in exponential loss function if

$$\eta(x) = 0,1$$

Exponential and Logistic Loss

- The two loss functions share the same probability model (up to a scaling factor)
- However, logistic regression changes the exponential sensitivity $e^{|p|}$ to $\ln\left(1+e^{|p|}\right)$
- The logistic regression loss behave better than the exponential loss.

Analysis of Hinge Loss

- $f_\phi^*(\eta) = sign(2\eta - 1)$

$$\Delta Q(\eta, p) = \begin{cases} (p-1)(1-\eta) + (1 - sign(2\eta-1))|2\eta - 1| & p \geq 1 \\ (p - sign(2\eta - 1))|2\eta - 1| & p \in [-1,1] \\ (p+1)\eta + (1 + sign(2\eta - 1))|2\eta - 1| & p \leq -1 \end{cases}$$

Hinge Loss: interpretation

- If $\eta(x)$ is close to 0.5: $f(x) - Truncation(f(x))$ is small
- Otherwise:

  if $|f(x)| \le 1$: $|f(x) - sign(2\eta - 1)|$ is small

  otherwise: $|f(x) - sign(2\eta(x) - 1)||2\eta(x) - 1 - sign(f(x))|$ is small

Maximum Margin

- Roughly speaking, if $\eta(x)$ is not close to 0.5, we require

$$f(x) \approx sign(2\eta(x) - 1)$$

- But allow

$$f(x) > 1 \qquad when \quad \eta(x) \approx 1$$

$$f(x) < -1 \qquad when \quad \eta(x) \approx 0$$

- This corresponds to the margin argument which motivated SVM and has been used to explain the effectiveness of boosting.

Comments on Margin Analysis

- Tong Zhang, 2004
- The margin idea is mostly useful in nearly separable cases ($\eta(x)$ is close to 0 or 1)
- It is not very useful if $\eta(x)(1-\eta(x))$ is not small.

Disadvantage

- The predictor computed by SVM does not carry any reliable probability information.
- By looking at the output of a SVM classifier at any given point, it is difficult to tell how confident the prediction is. Generally, such confidence information is often extremely valuable in practical applications.

Another Viewpoint

- The knowledge of density functions would allow us to solve whatever problems that can be solved on the basis of available data;
- Vapnik's principle: *never to solve a problem that is more general than you actually need to solve.*
- One should try to avoid estimating any density when solving a particular learning problem.

The Outline

- Some comments on margin-based generalization bound
- The convex upper bounds of classification loss function
- The statistical behavior of different loss functions
- Application in cost-sensitive classification problems

Cost-sensitive Problems

- Cost-sensitive classification considers different costs of each misclassified example.
- An cost-sensitive classification technique takes the cost of samples into consideration during model building and generates a model that has the lowest cost.

Recent Development: ICML

- H. Masnadi-Shirazi and N. Vasconcelos. Asymmetric boosting. *ICML*. 2007. They obtain a natural cost-sensitive AdaBoost, which is based on the statistical interpretation, i.e., minimizing the cost-sensitive loss by gradient descent in function space.

- They use loss function $E_x\left(\sum_{y=1,-1} e^{-c(x,y)yf(x)}\right)$

Recent Development: PR

- Y. Sun et al. Cost-sensitive boosting for classification of imbalanced data. Pattern Recognition 2007.
- Loss functions: $\sum_{y=1,-1} e^{-c(x,y)yf(x)}$ $\sum_{y=1,-1} c(x,y)e^{-yf(x)}$ $\sum_{y=1,-1} c(x,y)e^{-c(x,y)yf(x)}$
- The AdaBoost which optimizes $\sum_{y=1,-1} c(x,y)e^{-yf(x)}$ has furnished better results in most experiments.

Our Analysis

- A general AdaBoost framework for binary cost-sensitive classification is intuitively established.
- Theoretical analysis indicates that the $\sum_{y=1,-1} c(x,y)e^{-yf(x)}$ based AdaBoost has better performance.
- Further, the modified LogitBoost has better performance.

Summary

- Many of the classification algorithms developed in the machine learning literature, including SVM and boosting, can be viewed as optimization methods that minimize a convex surrogate of the 0-1 loss function.

- The convexity makes these algorithms computationally efficient. However, the use of a surrogate function has many significant statistical consequences.

Other Significant Issues

- Consistency of a Function Class
- Convergence rate

- **Thanks**