Statistical Behavior and Consistency of Classification Methods Based on Convex Risk Minimization by Tong Zhang

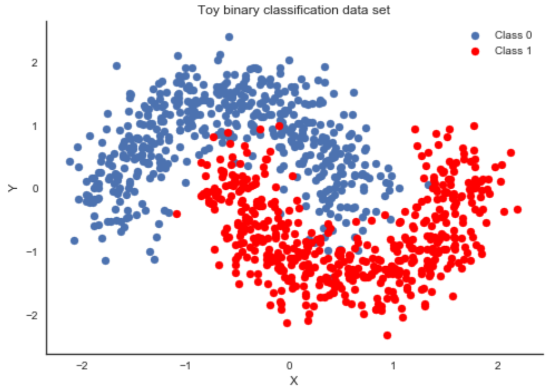
Philip Gigliotti, Jared Wilson, Evan Hart and Quan Wen

MAT 593 Practical Methods in Machine Learning

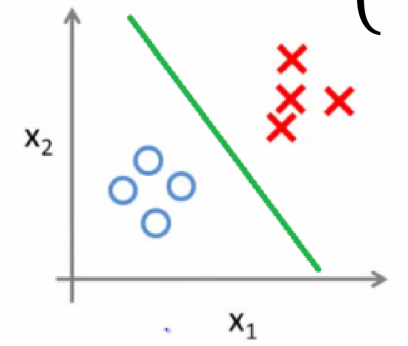
UNIVERSITY AT ALBANY
State University of New York

What is a Binary Classifier?

- Definition:
  I. Given data $\{x_i, y_i\}_{i=1}^{n}$ where:
  $x_i \in R^d$ are the features
  $y_i \in \{1, -1\}$ are the targets or class of the input x_i.
  II. Define a function $\mathcal{F}: R^d \to R$
  III. Now we consider the prediction rule $\hat{y}_i = \begin{cases} 1, & if\ f(x_i) \geq 0 \\ -1, & otherwise \end{cases}$



https://docs.pymc.io/notebooks/bayesian_neural_network_advi.html



https://utkuufuk.github.io/2018/06/03/one-vs-all-classification/

Philip Gigliotti, Jared Wilson, Evan Hart and Quan Wen
MAT 593 Practical Methods in Machine Learning

What is Classification Error (Zero-One Loss)?

- Let $\widehat{y_i}$ = the predicted class y_i predicted by the classifier $f(x_i)$
- Classification Error (zero-one loss):

$$I(f(x_i), y_i) = \begin{cases} 1, & if\ \hat{y}_i \neq y_i \\ 0, & if\ \hat{y}_i = y_i \end{cases}$$

- The classifier seeks to minimize zero-one loss given by:

$$L(f(.)) = E_{X,Y} I(f(x), y)$$

- **PROBLEM:** $I(f(x), y)$ is not convex! Thus classifier algorithms seek a convex approximation of this problem.

Convex Approximation of Error Minimization

- Minimization of zero-one loss is very difficult!
- ML algorithms rely on a convex loss function ϕ which approximates zero-one loss:

$$\phi(f(x), y)$$

- Common classifier algorithms minimize the zero-one loss approximation:

$$Q\big(f(.)\big) = E_{X,Y}\phi(f(x), y)$$

- **How good is this approximation?**
- **How close is the resulting error rate to the lowest possible error rate we could expect?**
- This paper explores these questions and suggests that many common classifiers approach this lower bound.

Setting Goals for Classification Error

- **Bayes Error Rate:** The lowest error rate possible for classification of a given random outcome. Based on the Bayes decision rule for classification.
- Given the conditional probability of class 1:

$$\eta(x) = p(Y = 1|X = x)$$

- The Bayes decision rule is a classifier $f^*(x) = \begin{cases} 1, & if\ \eta(x) \geq .5 \\ -1, & otherwise \end{cases}$
- For any function f, $L(f(.)) - L(f^*(.)) \geq 0$
- Thus $L(f^*(.)) = inf_f L(f(.))$

Bayes Error Rate and Consistency

- Classifiers that achieve the Bayes Error rate:
  - Logistic regression: We know the MLE, thus we know it can achieve Bayes Error Rate under certain conditions.
- Other common classifiers meet or achieve similar classification error rates to logistic regression. *Is it possible that they also achieve Bayes Error Rate?*
- A classifier is said to be **consistent** if it's classification error rate approaches the Bayes Error as the number of empirical observations approaches infinity.
- This paper explores 5 common binary classifiers, showing that they are consistent, and that their error converges to the Bayes Error rate.

Philip Gigliotti, Jared Wilson, Evan Hart and Quan Wen
MAT 593 Practical Methods in Machine Learning

Preface to Main Result

- Zhang proves the classifiers are consistent intuitively:
  - Introduce the concept of approximation error.
  - Recall the convex approximation of zero-one loss:

$$Q\big(f(.)\big) = E_{X,Y}\phi(f(x), y)$$

  - Establish the convex minimizer of the function, which may not be uniquely determined:

$$f_\phi^*(.) = arg\min_{f \in R} Q\big(f(.)\big)$$

  - Determine the minimal error rate possible through convex approximation:

$$Q^* = Q(f_\phi^*(.))$$

  - Define the approximation error of a given estimator as the difference between its error rate and the quantity Q*:

$$\Delta Q\big(f(.)\big) = Q\big(f(.)\big) - Q^*$$

Main Result

- Zhang proves the consistency of binary estimators based on minimization of the convex approximation for zero-one loss.
- Demonstrate that the difference between the error rate of a given estimator and the Bayes error rate (L*) is bounded by approximation error:

$$If\ |.5 - \eta|^s \leq c^s \Delta Q(\eta, 0)$$
$$then\ L(f(x)) - L^* \leq 2c\Delta Q\big(f(.)\big)^{1/s}$$

- Calculate the quantity $\Delta Q\big(f(.)\big)$ for 5 common estimators and show that for each estimator $\hat{f}(.)$:

$$\lim_{n\to\infty} \Delta Q\big(\hat{f}(.)\big) = 0$$

- **RESULT:** The error rates of these 5 common loss functions converge to the Bayes Error Rate.

Demonstration

- To show the following condition for each estimator:

  1) If $|.5-\eta| \leq c^s \Delta Q(\eta, 0)$
  2) then $L(f(x)) - L^* \leq 2c\Delta Q\left(f(.)\right)^{1/s}$

  we calculate $\Delta Q(\eta, f)$ and show that 1) holds.

- $\Delta Q(\eta, f)$ can be calculated with the formula:

$$\Delta Q(\eta, p) = \eta d_\phi\left(f_\phi^*(\eta), p\right) + (1-\eta) d_\phi\left(-f_\phi^*(\eta), -p\right)$$

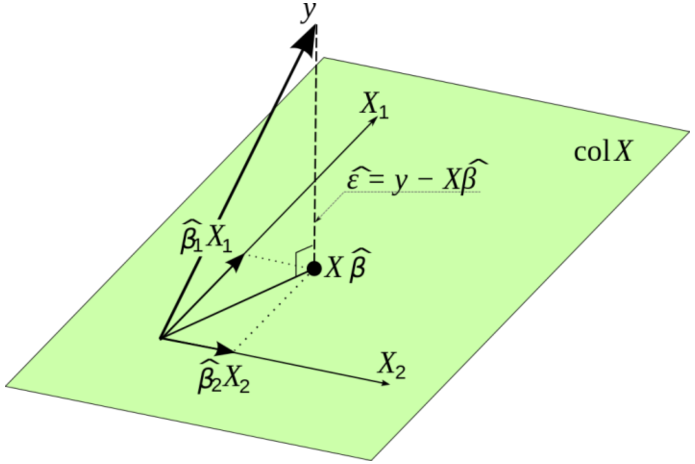
  where d_ϕ is Bregman divergence, calculated:

$$d_\phi(f_1, f_2) = (f_2) - \phi(f_1) - \phi'(f_2)(f_2 - f_1)$$

- To demonstrate this property, we need to calculate f_ϕ^*, $d_\phi(f_1, f_2)$, and $\Delta Q(\eta, p)$.

Example: OLS regression

- To find f_ϕ^*, we minimize the $Q(f(.))$ for the OLS loss function $\phi(v) = (1-v)^2$.
- $Q(f(.))$ can be re-formulated $Q(\eta, f) = \eta\phi(f) + (1-\eta)\phi(-f)$
- We solve the problem:
  $f_\phi^*(\eta) = \min_f Q(\eta, f) = \min_f \eta(1-f)^2 + (1-\eta)(1+f)^2 = 2\eta - 1$



https://en.wikipedia.org/wiki/Ordinary_least_squares

OLS Bregman Divergence and $\Delta Q(\eta, p)$

- $d_\phi(f_1, f_2) = \phi(f_2) - \phi(f_1) - \phi'(f_2)(f_2 - f_1)$
  $= (1 - f_2)^2 - (1 - f_1)^2 - 2(f_2 - 1)(f_2 - f_1)$
  $= (f_2 - f_1)^2$
- $\Delta Q(\eta, p) = \eta d_\phi\left(f_\phi^*(\eta), p\right) + (1 - \eta) d_\phi\left(-f_\phi^*(\eta), -p\right)$
  $= \eta(p - (2\eta - 1))^2 + (1 - \eta)\left((2\eta - 1) + p\right)^2$
  $= (2\eta - 1 - p)^2$

Boundedness of OLS Result

- If $\Delta Q(\eta, p) = (2\eta - 1 - p)^2$, clearly $\Delta Q(\eta, 0) = (2\eta - 1)^2$
- Rearranging, we get:

$$|.5 - \eta|^2 \leq .5^2\, \Delta Q(\eta, 0)$$

which fits the condition for our key result

$$|.5 - \eta|^s \leq c^s \Delta Q(\eta, 0)$$

where c = .5 and s = 2.

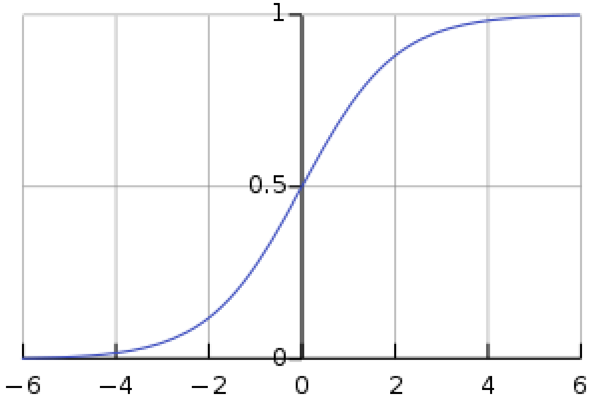
- Therefore, $L(f(x)) - L^* \leq 2c\Delta Q\left(f(.)\right)^{1/s}$

and since $\lim_{n\to\infty} \Delta Q\left(\hat{f}(.)\right) = 0$ for OLS, OLS is consistent.

Properties of OLS

- The function $p = 0$ minimizes $\Delta Q(\eta, p)$ when $\eta = .5$, since $f_{\phi}^{*} = 2\eta - 1$
- Using our classification decision rule, the function $f = 0$ estimates a probability of .5 that each observation is class 1.
- If $|.5 - \eta|^2 \leq .5^2\, \Delta Q(\eta, 0)$, we see that minimizing $\Delta Q(\eta, p)$ also minimizes the squared absolute difference of the predictor and in class probability η.
- Also, we see $(f(x) + 1)/2$ approximates the true in class probability.
- Therefore OLS provides a reliable estimate of the true in class probability.
- This provides us with a reliable estimate of how confident we can be in our prediction.

Example: Logistic Regression



https://en.wikipedia.org/wiki/Logistic_regression

- $f_\phi^*(\eta) = \ln \frac{\eta}{1-\eta}$
- $d_\phi(f_1, f_2) = -\eta \ln \eta - (1-\eta) \ln(1-\eta)$
- $\Delta Q(\eta, p) = \frac{1}{2\eta'(1-\eta')}(\eta - \bar{\eta}) \geq 2(\eta - \bar{\eta})^2$ where

  $\bar{\eta} = {f_\phi^*}^{-1}(p)$ and $\exists\, \eta'$, the Taylor expansion between $\bar{\eta}$ and η.

- Therefore:

$$\Delta Q(\eta, 0) \geq 2(\eta - .5)^2 \; or$$

$$|.5 - \eta|^2 \leq .5\Delta Q(\eta, 0) = 2^{-\frac{1}{2}^2} \Delta Q(\eta, 0)$$

with s = 2, c = $2^{-1/2}$

Properties of Logistic Regression

- The logistic transform $1/(1+e^{-f(x)})$ of f(x) approximates the true conditional in-class probability.
- Therefore, logistic regression is a maximum likelihood estimator.
- However, when the conditional in class probability is very close to 0 or 1, |f(x)| must be very large to approximate such a value given the logistic transform.
- Accordingly, logistic regression is poorly suited to predicting rare events.

Example: Support Vector Machines

- $f_\phi^*(\eta) = sign(2\eta - 1)$
- Cannot calculate Bregman divergence since ϕ' is not uniquely defined.
- Calculated directly,
  $\Delta Q(\eta, p) = \eta \max(0, 1 - p) + (1 - \eta) \max(0, 1 + p) - 1 + |2\eta - 1|$
- Therefore:

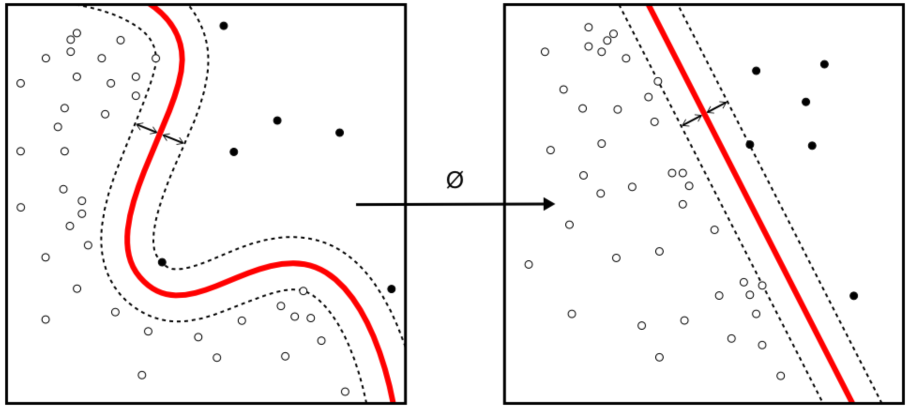
$$\Delta Q(\eta, 0) = \eta + (1 - \eta)1 + |2\eta - 1| = |2\eta - 1|$$

so

$$|\eta - .5| = .5\Delta Q(\eta, 0) \text{ with s= 1 and c= .5}$$

Properties of Support Vector Machines

- SVM delivers reliable predictions of class, and allows for conditions when $\eta \approx 1$ or $\eta \approx 0$
- However, even when $\eta(1-\eta)$ is not close to zero, f(x) clusters at 1 and -1.
- Therefore, SVM does not give reliable information about the confidence level of a prediction.



https://en.wikipedia.org/wiki/Support-vector_machine

Criticism

- Convexity, Classification and Risk Bounds by Bartlett Jordan, McAuliffe
  - “Need to find general quantitative relationships between approximation and estimation errors associated with Φ and those associated with 0-1 loss.”
  - Zhang presents several examples of these relationships.
  - This paper aims to simplify and extend Zhang’s results.
  - Similar comparison theorems to parts 1 and 3b of Theorem 1
  - However, authors’ conclusions hold under weaker conditions than those assumed by Zhang.
  - Pgs. 142-143

Criticism

- Convexity, Classification, and Risk Bounds by Bartlett, Jordan, and McAuliffe
  - Difficult of pattern classification is related to behavior of conditional in-class probability $\eta(X)$.
  - In practical problems, it is reasonable to assume for most X that $\eta(X)$ is not too close to ½
  - Tsybakov (2001) introduced formulation of this assumption
    - Under assumption of low noise, risk converges quickly to the minimum over the class
    - If minimum is nonzero, we expect a convergence rate as fast as $1/n$
  - Authors show that minimizing empirical Φ-risk leads to fast convergence rates under Tsybakov's assumption
  - If Φ is uniformly convex, empirical Φ-risk converges quickly to Φ-risk
  - Noise assumption allows improvement in relationship between excess Φ-risk and excess risk

Criticism

- Theory of Classification: A Survey of Some Recent Advances by Boucheron, Bousquet, and Lugosi
  - Excess misclassification error L(f)-L* is related to excess loss A(f)-A*
  - According to [27] (Bartlett, Jordan, McAuliffe), the above lemma may be improved under Mammen-Tsybakov noise conditions to yield

$$L(f) - L(f^*) \leq \left(\frac{2^s c}{\beta^{1-s}}(A(f) - A^*)\right)^{1/(s-s\alpha+\alpha)}.$$

  - The refined bounds may be carried over to analysis of classification rules based on the empirical minimization of a convex cost functional
  - The bounds are tighter and thus more precise, resulting in faster rates of convergence