Empirical risk minimization for classification

김지수 (Jisu KIM)

통계적 기계학습(Statistical Machine Learning), 2024 1st semester

The lecture note is a minor modification of the lecture notes from Prof. Yongdai Kim's "Statistical Machine Learning".

1 Review

1.1 Basic Model for Supervised Learning

- Input(입력) / Covariate(설명 변수) : $x \in \mathbb{R}^d$, so $x = (x_1, \ldots, x_d)$.
- Output(출력) / Response(반응 변수): $y \in \mathcal{Y}$. If y is categorical, then supervised learning is "classification", and if y is continuous, then supervised learning is "regression".
- Model(모형) :

  $$y \approx f(x).$$

  If we include the error ϵ to the model, then it can be also written as

  $$y = \phi(f(x), \epsilon).$$

  For many cases, we assume additive noise, so

  $$y = f(x) + \epsilon.$$

- Assumption(가정): f belongs to a family of functions $\mathcal{M}$. This is the assumption of a model: a model can be still used when the corresponding assumption is not satisfied in your data.
- Loss function(손실 함수): $\ell(y, a)$. A loss function measures the difference between estimated and true values for an instance of data.
- Training data(학습 자료): $\mathcal{T} = \{(y_i, x_i), i = 1, \ldots, n\}$, where (y_i, x_i) is a sample from a probability distribution P_i. For many cases we assume i.i.d., or x_i's are fixed and y_i's are i.i.d..
- Goal(목적): we want to find f that minimizes the expected prediction error,

  $$f^0 = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(Y,X)\sim P} \left[\ell(Y, f(X))\right].$$

  Here, $\mathcal{F}$ can be different from $\mathcal{M}$; $\mathcal{F}$ can be smaller then $\mathcal{M}$.
- Prediction model(예측 모형): f^0 is unknown, so we estimate f^0 by $\hat{f}$ using data. For many cases we minimizes on the empirical prediction error, that is taking the expectation on the empirical distribution $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{(Y_i,X_i)}$.

  $$\hat{f} = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{P_n} \left[\ell(Y, f(X))\right] = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(X_i)).$$

- Prediction(예측): if $\hat{f}$ is a predicted function, and x is a new input, then we predict unknown y by $\hat{f}(x)$.

1.2 Bayes Classifier

For the classification so that $y \in \mathcal{Y} = \{1, \ldots, J\}$, the Expected Prediction Error (EPE) is

$$\begin{aligned} EPE(f) &= \mathbb{E}_{(Y,X)\sim P}\left[\ell(Y, f(X))\right] = \mathbb{E}\left[I(Y \neq f(X))\right] \\ &= \mathbb{E}_X \left[\sum_{j=1}^{J} \ell(j, f(X)) P(Y = j|X)\right]. \end{aligned}$$

Hence $EPE(f)$ is minimized when

$$f^0(x) = \arg\min_{k=1,\ldots,J} \sum_{j=1}^{J} \ell(j,k) P(Y = j|X = x).$$

For the classification, the most common loss is the 0-1 loss

$$\ell(y, a) = I(y \neq a).$$

And the optimal prediction function is

$$f^0(x) = \arg\max_{j=1,\ldots,J} P(Y = j|X = x).$$

This optimal classifier is called Bayes rule / Bayes classifier (베이즈분류 / 베이즈모형), and the error rate is called the Bayes risk (베이즈위험).

2 Introduction

The objective of statistical learning is to find

$$f^0 = \arg\min_{f\in\mathcal{F}} \mathbb{E}_{(Y,X)\sim P}\left[\ell(Y, f(X))\right].$$

The ERM principle is to estimate f^0 by $\hat{f}$ given by

$$\hat{f} = \arg\min_{f\in\mathcal{F}_n} \mathbb{E}_{P_n}\left[\ell(Y, f(X))\right] = \arg\min_{f\in\mathcal{F}_n} \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)).$$

where $\mathcal{F}_n$ is a set of functions such that $\mathcal{F}_n \to \mathcal{F}$ in some sense. That is, we estimate f^0 by minimizing the empirical risk.

Two components in supervised learning:

- Loss l
- A class of functions $\mathcal{F}_n$ (and $\mathcal{F}$).

Examples of $\mathcal{F}_n$:

- Linear functions
- Linear combinations of decision trees
- Reproducing kernel Hilbert space

This lecture considers the choice of the loss function. For two class problems ($y \in \{-1, 1\}$), the gold standard loss function is 0-1 loss $l(y, f) = I(y \neq f)$. However, minimizing the empirical risk with respect to the 0-1 loss is computationally impossible (NP complete) because the loss is discontinuous.

To resolve this problem, we use surrogated losses. The main theme of this chapter is to introduce various surrogated losses and compare them.

3 ERM with surrogated loss for classification

The Idea of using surrogated losses is as follows:

- $\mathcal{F}$: a set of functions $f : \mathbb{R}^p \to \mathbb{R}$
- Find $\hat{f} = \arg\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i))$ for some easy loss l.
- Make the classifier by $Sign(\hat{f}(x))$.

Examples of surrogated losses are:

- Exponential (Boosting): $l(y, a) = \exp(-ya)$
- Square (NN) : $l(y, a) = (y - a)^2$
- Logistic (MART) : $l(y, a) = \log(1 + \exp(-ya))$
- Hinge loss (SVM) : $l(y, a) = (1 - ya)_+$

These surrogates losses are

- Upper bound of 0-1 loss
- Convex
- Strictly convex except Hinge loss
- Differentiable except Hinge loss

Then we can ask following questions: what properties are necessary for a given loss to be a good surrogated loss for classification? (Fisher consistency), and which loss is better? (Convergence rate)

4 Fisher Consistency

- Under certain regularity conditions, we expect that $\hat{f}$ converges to f^0 where

$$f^0 = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(Y,X)\sim P} \left[\ell(Y, f(X))\right].$$

- Fisher consistency means that $sign(f^0(x))$ is the Bayes classifier.
- Then, we can expect that $sign(\hat{f}(x))$ is an (good?) estimator of the Bayes classifier.
- Now, the question is what properties are necessary for Fisher consistency.

For necessary conditions, we consider a margin based loss function:

- Suppose $y \in \{-1, 1\}$.
- A margin based loss function is when $l(y, f)$ be a function of the margin yf.
- The four losses (exponential, square, logistic and hinge losses) are all margin based losses.
- From now on, we only consider margin based loss functions.
- We use $l(yf)$ instead of $l(y, f)$.

Sufficient conditions for Fisher consistency is by Lin (2001):

- Suppose
  - $l(z) < l(-z)$ for all $z > 0$
  - $l^{'}(0) \neq 0$ exists.
- Then l satisfies the Fisher consistency.

- Note that all the four losses satisfies the Fisher consistency.
- More refined studies have been done by Bartlett et al. (2006).

We can show the followings:

- Exponential loss: $f^0(x) = \frac{1}{2} \log P(y = 1|x)/P(y = -1|x)$
- Squared error loss: $f^0(x) = P(y = 1|x) - 1/2$
- Logistic loss: $f^0(x) = \log P(y = 1|x)/P(y = -1|x)$
- Hinge loss: $f^0(x) = sign(P(y = 1|x) - 1/2)$. (see Lin (2002) and Bartlett et al. (2006) for proof).

5 Comparison of surrogated loss functions

- There are many surrogated losses which satisfy the Fisher consistency.
- Now, the question is which loss is better?
- The convergence rate plays a key role for answering this problem.
- This is one of the hottest topics of theoretical statistics.

We first review general facts for convergence rate:

- Let $\hat{f}_n$ be any estimator with sample size n and f^* is the target function.
- Suppose $\hat{f}_n$ and f^* are smooth.
- Then, under regularity conditions, it is well known that $\|\hat{f}_n - f^*\| \geq O_p(n^{-1/2})$.
- We can think of $\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} N(0, \sigma^2)$.
- Nonparametric estimator cannot have faster rate of convergence than parametric one.

In classification, we are interested in the convergence of risk rather than the convergence of classifier.

- That is, we are to study $R(\hat{f}) - R(f^0)$ where $R(f) = \mathbb{E}(I(Y \neq sign(f(X)))$.
- Marron (1983) showed that if we construct a classifier by estimating the density optimally (i.e. $\sqrt{n}$ rate), the optimal convergence rate of $R(\hat{f}) - R(f^0)$ is $\sqrt{n}$.
- Later, Mammen and Tsybakov (1999) showed that the convergence rate can be faster than $\sqrt{n}$ (up to n) if we can directly estimate the classifier (i.e. decision boundary).
- This result is reasonable since the complexity of the decision boundary is much smaller than the complexity of the density (or probability).
- That is, theoretically, classification is simpler than regression. However, computationally, classification is harder since we don't know how to estimate the decision boundary directly.
- We can expect to have a faster rate of convergence by choosing the surrogated loss appropriately.

We have the following heuristics for the choice of loss functions for classification.

- Theoretically, using the 0-1 loss in the ERM framework gives a best result (since we can estimate the decision boundary directly). However, we cannot do that due to computational difficulty.
- One conjecture is that if the surrogated loss is closer to the 0-1 loss, it has a better accuracy.
- Lin (2002) proved that the Hinge loss is a tight upper bound of the convex surrogated loss of the 0-1 loss. Hence, we expect that the SVM works better.

- Another heuristic argument of the Hinge loss is that the population risk minimizer of the hinge loss is a decision boundary not a probability.
- Shen (2003) suggested the ψ-loss, which is non-convex and closer to the 0-1 loss than the hinge loss. Also, he proved in some situations that the faster convergence rate is achieved.
- Steinwart and Scovel (2007) proved that SVM achieves fast convergence rate with Gaussian kernels.
- Tail behavior of $l(z)$ when z is small is important for convergence rate.
- Smaller the values of $l(z)$ when z is small, the better the accuracy.
- For the 0-1 loss, $\sup_{z<0} l(z) = 1$.
- For hinge and logistic losses, $l(z) = O(z)$ as $z \to \infty$.
- While $l(z) = O(\exp(z))$ and $l(z) = O(z^2)$ for the exponential and squared error losses.
- This partly explains that the MART (or logit boost) is better than AdaBoost.

6 References

- Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. Journal of the American Statistical Association, 101(473):138-156, 2006
- Lin, Y. (2001) A Note on Margin-based Loss Functions in Classification . Statistics and Probability Letters, **68**, 73-82.
- Lin, Y. (2002) Support Vector Machines and the Bayes Rule in Classification . Data Mining and Knowledge Discovery . 6. 259-275.
- Marron, J.S. (1983). Optimal rates of convergence to Bayes risk in nonparametric discrimination. *Ann. Statisti.*, **11**, 1142-1155.
- Mammen, E. and Tysbakov, A.B. (1999). Smooth discrimination analysis. *Ann. Statist.*, **27**, 1808-1829.
- Shen, X., Tseng, G., Zhnag, X. and Wong, W. (2003). On ψ-learning. *Journal of the American Statistical Association*, **98**, 724-734.
- Steiwart, I. and Scovel, C. (2007). Fast rates for support vector machines using Gaussian kernels. *Ann. Stat.* **35**, 575-607.