Risk and Loss Functions

Yong Wang
Dongqing Zhang

Friday, February 6, 2004

Outline

- Lost functions
- Test error & expected risk
- A statistical perspective
- Robust estimators
- Summary

Lost Functions: Overview

- Motivation: Determine a criterion according to which we will assess the quality of an target estimation based on observations during the learning
  - Non-trivial: the options are plenteous
- Definition:
  - The mapping c of the triplex (x, y, f(x)) into [0, infinite) with c(x, y, y) = 0
  - The minimum of the loss is 0 and obtainable, at least for a given x, y
  - In practice: the incurred loss is not always the quantity that we will attempt to minimize
    - Feasibility, confidence level consideration, …

Some Examples

- In classification
  - Misclassification error

$$c(x, y, f(x)) = \begin{cases} 0 & \text{if } y = f(x) \\ 1 & \text{otherwise} \end{cases}$$

  - Input-dependent loss

$$c(x, y, f(x)) = \max(0, 1 - yf(x)) = \begin{cases} 0 & \text{if } yf(x) \geq 1, \\ 1 - yf(x) & \text{otherwise.} \end{cases}$$

  - Asymmetric loss
  - Soft margin loss

$$c(x, y, f(x)) = \begin{cases} 0 & \text{if } y = f(x) \\ \tilde{c}(x) & \text{otherwise} \end{cases}$$

  - Logistic loss

$$c(x, y, f(x)) = \ln\left(1 + \exp\left(-yf(x)\right)\right)$$

Some Examples (cont.)

- In regression
  - Usually associated with the degree of the difference

$$c(x, y, f(x)) = \tilde{c}(f(x) - y)$$

  - Squared error

$$c(x, y, f(x)) = (f(x) - y)^2$$

  - _-insensitive loss

$$\tilde{c}(\xi) = \max(|\xi| - \varepsilon, 0) =: |\xi|_\varepsilon$$

- Criterion in practice
  - Cheap to compute
  - Small number of discontinuities in the first derivative
  - Convex to ensure the uniqueness of the solution
  - Outlier resistance

Test Error & Expected Risk

- Motivation: Given errors penalized on specific instances (x, y, f(x), how to combine these penalties to assess a particular estimation f
- Definition of test error

$$R_{\text{test}}[f] := \frac{1}{m'} \sum_{i=1}^{m'} \int_{\mathcal{Y}} c(x'_i, y, f(x'_i)) dP(y|x'_i)$$

  - Hard to resolve
- Definition of expected risk

$$R[f] := \mathbf{E}\left[R_{\text{test}}[f]\right] = \mathbf{E}\left[c(x, y, f(x))\right] = \int_{\mathcal{X}\times\mathcal{Y}} c(x, y, f(x)) dP(x, y)$$

  - Situation is not becoming better: P(x, y) is unknown
  - Simplification: empirical estimation using training patterns

Approximations

- Assumptions
  - The existence of a underlying probability distribution P(x, y) governing the data generation
  - Data (x, y) are drawn i.i.d. from P(x, y)
  - pdf p(x,y) exists
- Empirical density

$$p_{\text{emp}}(x, y) := \frac{1}{m} \sum_{i=1}^{m} \delta_{x_i}(x)\delta_{y_i}(y)$$

  - Lead to a quantity "reasonably close" to the expected risk
- Empirical risk

$$R_{\text{emp}}[f] := \int_{\mathcal{X}\times\mathcal{Y}} c(x, y, f(x)) p_{\text{emp}}(x, y) dx dy = \frac{1}{m} \sum_{i=1}^{m} c(x_i, y_i, f(x_i))$$

  - Risk of rising ill-posed problems
  - Overfitting

Ill-posed Problem: Example

- Address a regression problem using quadratic loss function
- Dealing with a linear class of functions

$$\mathfrak{F} := \left\{ f \,\middle|\, f(x) = \sum_{i=1}^{n} \alpha_i f_i(x) \text{ with } \alpha_i \in \mathbb{R} \right\}$$

$$\underset{f \in \mathfrak{F}}{\text{minimize}}\, R_{\text{emp}}[f] = \underset{\boldsymbol{\alpha} \in \mathbb{R}^n}{\text{minimize}}\, \frac{1}{m} \sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{n} \alpha_j f_j(x_i) \right)^2$$

$$F^\top \mathbf{y} = F^\top F \boldsymbol{\alpha} \quad \Longrightarrow \quad \boldsymbol{\alpha} = \left(F^\top F\right)^{-1} F^\top \mathbf{y} \quad \text{where} \quad F_{ij} := f_i(x_j)$$

- If we have more basis functions f_i than observations, there will be a subspace of solutions

A Statistical Perspective

- For a given observation and its estimation, besides what risk we can expect for it, we may be interested in which probability the corresponding loss is going to occur
- Need to compute p(y|x)
  - Should be aware there are two approximations
    - Model the density p firstly
    - Compute a minimum of the expected risk
  - This could lead to inferior or at least not easily predictable results
    - Additional approximation steps might make the estimates worse

Maximum Likelihood Estimation

- Likelihood

$$p(\{x_1,\ldots,x_m\},\{y_1,\ldots,y_m\}|f) = \prod_{i=1}^{m} p(x_i, y_i|f) = \prod_{i=1}^{m} p(y_i|x_i, f)p(x_i)$$

- Log-Likelihood

$$\mathcal{L}[f] := \sum_{i=1}^{m} -\ln p(y_i|x_i, f)$$

- Minimization of Log-likelihood coincides with empirical risk if the loss function c is chosen according to $c(x, y, f(x)) = -\ln p(y|x, f)$
  - For regression: $c(x, y, f(x)) = -\ln p_{\xi}(y - f(x))$
    - _ is the additive noise to f(x) with density *p*_
  - For classification: $c(x, y, f(x)) = -\ln \mathrm{P}(y|f(x))$

Density Modeling

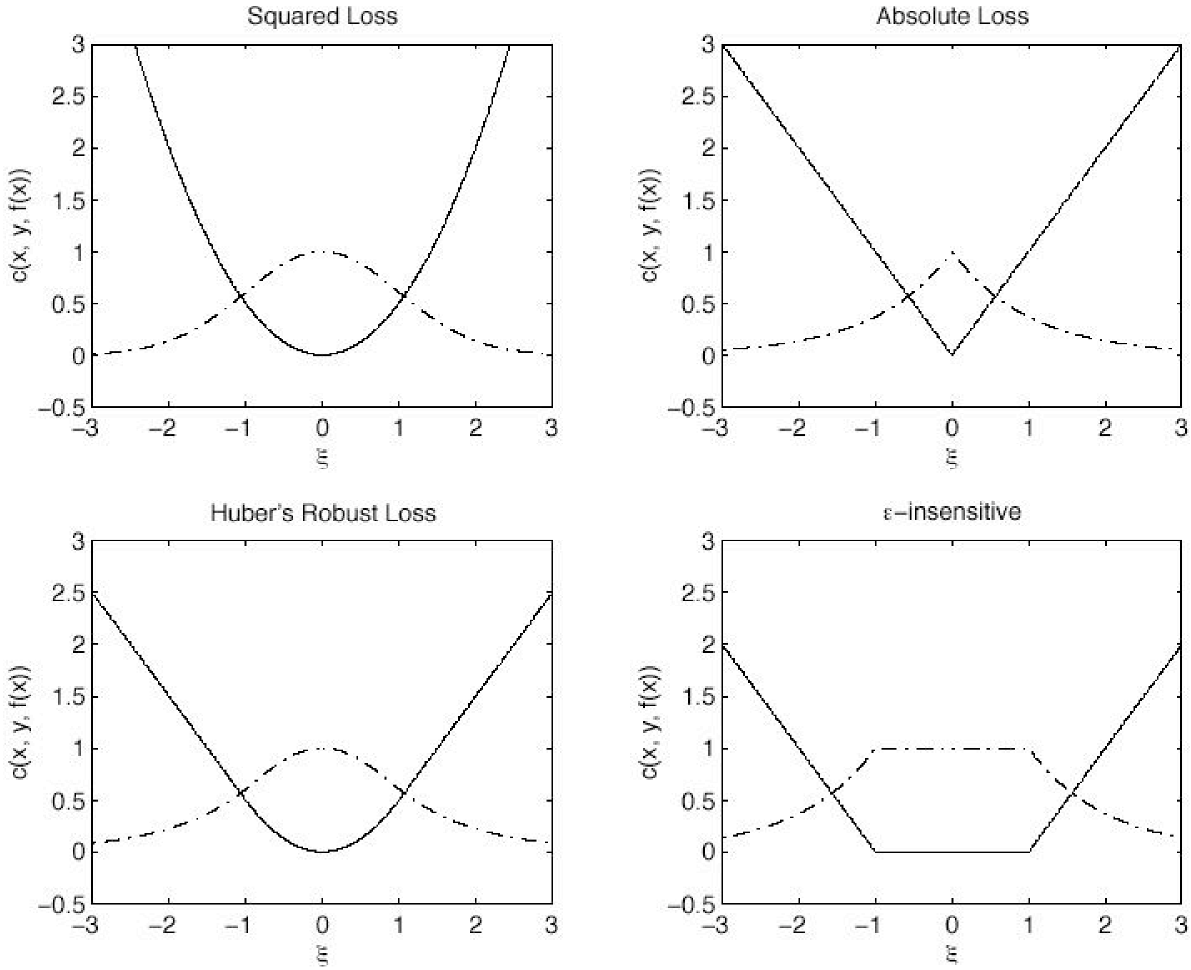
- Possible models
  - Logistic transfer function
  - Probit model
  - Inverse complementary log-log model
  - Q: what's the policy to select a suitable model?
- For classification:Logistic model & loss function

$$P(y=1|x,f) := \frac{\exp(f(x))}{1+\exp(f(x))}$$

$$-\ln P(y=1|x,f) = \ln(1+\exp(-f(x)))$$

- For regression: see next page

Loss Functions & Density Models



_ density models for regression $y_i = f_{\text{true}}(x_i) + \xi_i$

Practical Consideration

- Loss functions resulting from a maximum likelihood reasoning might be non-convex
- Strong assumption: explicitly we know $P(y|x, f)$
- The minimization of log-likelihood depends on the class of functions
  - No better situation than by minimizing empirical risk
- Is the choice of loss function arbitrary?
  - Does there exist good means of assessing the performance of an estimator?
  - Solution: efficiency
    - How noisy an estimator is with respect to a reference estimator

Estimator

- Denote by P(y|_) a distribution of y depending on the parameters _ (might be a vector), and by Y={y_1, …, y_m} an m-sample drawn i.i.d. from P(y|_)
- Estimator $\hat{\theta}(Y)$ of the parameter based on Y

$$\mathbf{E}_\theta\left[\xi(y)\right] := \mathbf{E}_{\mathrm{P}(y|\theta)}\left[\xi(y)\right] = \int \xi(y) d\mathrm{P}(y|\theta)$$

- Unbiased assumption

$$\mathbf{E}_\theta\left[\hat{\theta}(Y)\right] = \theta$$

- The efficient way to compare unbiased estimators is to compute their variance
  - The smaller the variance, the lower the probability $\hat{\theta}(Y)$ will deviate from _
  - Use variance as a one-number performance measure

Fisher Information, etc.

- Score function

$$V_\theta(Y) := \partial_\theta \ln p(Y|\theta) = \partial_\theta \sum_{i=1}^{m} \ln p(y_i|\theta) = \sum_{i=1}^{m} \frac{\partial_\theta p(y_i|\theta)}{p(y_i|\theta)}$$

  - Indicating how much the data affect the choice of _
- Covariance of V_(Y) is called the Fisher information matrix *I*

$$\mathbf{E}_\theta\left[V_\theta(Y)\right] = \int p(Y|\theta)\partial_\theta \ln p(Y|\theta)dY = \partial_\theta \int p(Y|\theta)dY = \partial_\theta 1 = 0$$

$$I_{ij} := \mathbf{E}_\theta\left[\partial_{\theta_i} \ln p(Y|\theta) \cdot \partial_{\theta_j} \ln p(Y|\theta)\right]$$

- Covariance of the estimator $\hat{\theta}(Y)$

$$B_{ij} := \mathbf{E}_\theta\left[\left(\hat{\theta}_i - \mathbf{E}_\theta\left[\hat{\theta}_i\right]\right)\left(\hat{\theta}_j - \mathbf{E}_\theta\left[\hat{\theta}_j\right]\right)\right]$$

Cramer & Rao Boundary

- Any unbiased estimator $\hat{\theta}(Y)$ satisfies $\det IB \geq 1$
  - $\hat{\theta}(Y)$ deviates from _ by more than a certain amount
  - The definition of a one-number summary of the properties of an estimator, namely how closely the inequality is met
- Efficiency: $e := 1/\det IB$
  - The closer *e* is to 1, the lower the variance of the estimator $\hat{\theta}(Y)$.
  - For a special class of estimators, B and e can be computed efficiently

Efficiency

- Asymptotic variance

Theorem 3.13 (Murata, Yoshizawa, Amari [379, Lemma 3]) *Assume that* $\hat{\theta}$ *is defined by* $\hat{\theta}(Y) := \operatorname{argmin}_{\theta} d(Y,\theta)$ *and that* d *is a twice differentiable function in* θ. *Then asymptotically, for increasing sample size* $m \to \infty$, *the variance* B *is given by* $B = Q^{-1}GQ^{-1}$. *Here*

$$G_{ij} := \operatorname{cov}_{\theta}\left[\partial_{\theta_i} d(Y,\theta), \partial_{\theta_j} d(Y,\theta)\right] \text{ and} \tag{3.38}$$

$$Q_{ij} := \mathbf{E}_{\theta}\left[\partial^2_{\theta_i,\theta_j} d(Y,\theta)\right], \tag{3.39}$$

and therefore $e = (\det Q)^2/(\det IG)$.

- Maximum Likelihood(ML) is asymptotically efficient
  - I.e., e=1, as m -> infinite

$$\hat{\theta}(Y) := \operatorname*{argmax}_{\theta} \ln p(Y|\theta) = \operatorname*{argmin}_{\theta} \mathcal{L}[\theta]$$

ML In Reality: No Perfect

- ML is efficient "asymptotically"
  - For finite sample size, it is possible to do better other than ML estimation
- Practical considerations such as the goal of sparse decomposition (?) may lead to the choice of a non-optimal loss function
- We may not know the true density model P(y|_), which is required to define the ML estimator
  - Definitely we can guess
  - While a bad guess can lead to large errors
  - Solution: robust estimators

Robust Estimators

- Practical assumptions
  - A certain class of distributions from which P(Y) is chosen
  - Training and testing data are identically distributed
- Robust estimators are used to safeguard us against the cases where the above assumptions are not true
- Avoid a certain fraction _ of 'bad' observations (outliers) seriously affecting the quality of the estimate
  - The influence of individual patterns should be bounded from above

Robustness via Loss Functions

- Basic idea (Huber): take a loss function as provided by the ML framework, and modify it in such a way as to limit the influence of each individual patter
  - Achieved by providing an upper bound on the slope of -ln[p(Y|_)]
  - Examples
    - trimmed mean or median
    - _-insensitive loss function

Robust Loss Function Theorem

Theorem 3.15 (Robust Loss Functions (Huber [250])) *Let $\mathfrak{P}$ be a class of densities formed by*

$$\mathfrak{P} := \{p | p = (1-\varepsilon)p_0 + \varepsilon p_1\} \text{ where } \varepsilon \in (0,1) \text{ and } p_0 \text{ are known.} \tag{3.47}$$

Moreover assume that both p_0 and p_1 are symmetric with respect to the origin, their logarithms are twice continuously differentiable, $\ln p_0$ is convex and known, and p_1 is unknown. Then the density

$$\bar{p}(\theta) := (1-\varepsilon)\begin{cases} p_0(\theta) & \text{if } |\theta| \leq \theta_0 \\ p_0(\theta_0)e^{-k(|\theta|-\theta_0)} & \text{otherwise} \end{cases} \tag{3.48}$$

is robust in the sense that the maximum likelihood estimator corresponding to (3.48) has minimum variance with respect to the "worst" possible density $p_{\text{worst}} = (1-\varepsilon)p_0 + \varepsilon p_1$: it is a saddle point (located at p_{worst}) in terms of variance with respect to the true density $p \in \mathfrak{P}$ and the density $\bar{p} \in \mathfrak{P}$ used in estimating the location parameter. This means that no density p has larger variance than p_{worst} and that for $p = p_{\text{worst}}$ no estimator is better than the one where $\bar{p} = p_{\text{worst}}$, as used in the robust estimator.

Practice Consideration

- Even though a loss function defined in Theorem 3.15 is generally desirable, we may be less cautious, and use a different loss function for improved performance, when we have additional knowledge of the distribution
- Trimmed mean estimator (Remark 3.17)
  - Discards _ of the data: effectively all $\_{i}$ deviating from the mean by more than _ are ignored and the mean is adjusted
  - When _ -> 1, we recover the median estimator: all patterns but the median one are discarded (?)

Efficiency & _-Insensitive Loss Function

- Use efficiency theorem, the performance of _-insensitive loss function can be estimated when applied to different types of noise model
- Gaussian Noise
  - If the underlying noise model is Gaussian with variance _ and _-insensitive loss function is used, the most efficient estimator from this family is given by _=0.612_
- More general:

$$\varepsilon_{\text{opt}} = \sigma \operatorname*{argmin}_{\tau} \frac{1}{\left(p_{\text{std}}(-\tau) + p_{\text{std}}(\tau)\right)^2} \left(1 - \int_{-\tau}^{\tau} p_{\text{std}}(\tau')d\tau'\right)$$

  - _ has to be known in advance
  - Otherwise: adaptive loss functions

Adaptive Loss Functions

- In _-insensitive loss function case, adjust _ with a small enough _ and see the loss changes
- Idea: for a given p(y|_), determine the optimal value of _ by computing the corresponding fraction _ of patterns outside the interval [-_+_, _+_].
- _ is found by Theorem 3.21

$$\nu = 1 - \int_{-\varepsilon}^{\varepsilon} p_{\text{std}}(y)dy$$

  - Given the type of additive noise, we can determine the value of _ such that it yields the asymptotically efficient estimator
  - Case study: polynomial noise model

Summary

- Two complementary concepts as to how risk and loss functions should be designed
  - Data driven: uses the incurred loss as its principal guideline
    - Empirical risk
    - Expected risk
  - Idea of estimating the distribution which may generate the data
    - ML is conceptually rather similar to the notions of risk & loss
    - Evaluate the estimator performance using Cramer-Rao theorem
    - How loss functions adjust themselves to the amount of noise, achieving optimal performance
    - _-insensitive loss function is extensively discussed as case study