Chapter 3: Risk and Loss Functions

Jan Eriksson

jan.eriksson@hut.fi

Signal Processing Laboratory

Helsinki University of Technology

Outline

- Loss functions and risk
- Statistical estimators and concepts
- Robustness

3.1 Loss functions

- The function $c: \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ with the property $c(x, y, y) = 0$ for all $(x, y, y) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$ is called a loss function.
- In pattern classification context, the value $c(x, y, f(x))$ represents the loss incurred for the *pattern* x and *target* y with a *prediction* function $f: \mathcal{X} \to \mathcal{Y}$.
- Notice that the incurred loss is not necessarily the quantity we want to minimize!

3.1.1 Binary Classification, i.e. $y \in \{-1, 1\}$

- Simplest case involves counting the misclassifications, i.e. misclassification error is given by

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

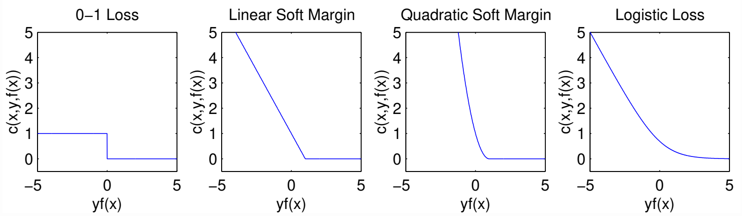
- Misclassification error does not distinguish between different types of errors (e.g. false positive or negative). Input-Dependent (Asymmetric) loss depends on input, i.e.

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

3.1.1 Binary Classification (2)

- Also confidence of the prediction may be of interest. Let $f(x)$ be now *real valued*. Then $\mathrm{sgn}(f(x))$ denotes the class label, $|f(x)|$ is the confidence (level) of the prediction, and the product $yf(x)$ can be used to assess the quality of the prediction in the loss functions.
- Such loss functions include soft margin loss function $c(x, y, f(x)) = \max(0, 1 - yf(x))$, quadratic soft margin loss $c(x, y, f(x)) = \max(0, 1 - yf(x))^2$, and logistic loss $c(x, y, f(x)) = \ln(1 + \exp(-yf(x)))$.

3.1.1 Binary Classification (3)



- Generalizations of confidence levels to Multi Class Discrimination (i.e. more than two classes) is not straightforward.

3.1.2 Regression, i.e. $y \in \mathbb{R}$

- When estimating the real valued quantities, the quality of misprediction is usually measured in terms of the size of the difference $y - f(x)$.
- Assuming location independence, the loss functions are thus functions the difference, i.e. $c(x, y, f(x)) = \tilde{c}(y - f(x))$.
- Such losses include squared loss $\tilde{c}(\xi) = \xi^2$ and ε-insensitive loss $\tilde{c}(\xi) = \max(|\xi| - \varepsilon, 0)$, which gives as a special case $(\varepsilon = 0)$ the ℓ_1 loss.
- Practical choice of the loss function is affected by properties like cheapness of computation, continuity (differentiation) properties, convexity, and robustness.

3.2.1 Test Error and Expected Risk (exact)

- Loss functions specify losses for specific (data) instances. Need for combining these local penalties (i.e. global measure for data).
- Assume that $\mathcal{X} \times \mathcal{Y}$ is governed by a probability distribution $P(x, y)$. Further, data is is supposed to be drawn i.i.d. Depending if test patterns are also given, two significantly different cases arises.

3.2.1 Test Error and Expected Risk (2)

- (Transduction Problem) Given the *training* data $\{x_1, \ldots, x_m\}$ along with *target values* $\{y_1, \ldots, y_m\}$ and the *test patterns* $\{x'_1, \ldots, x'_{m'}\}$ we want to minimize Test Error given by

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

i.e. we want to minimize the average loss *on the given test patterns*.

3.2.1 Test Error and Expected Risk (3)

- When the test patterns are not given, we want to minimize the average loss with respect to all test patterns, i.e. the Expected Risk

$$R[f] := \int_{\mathcal{X} \times \mathcal{Y}} c(x, y, f(x)) dP(x, y).$$

3.2.2 Approximations

- Unfortunately, the probability distribution is usually unknown, and all we have is the (training) data.
- In this case we can replace the probabilities with their *empirical* ones, which leads to Empirical Risk given by

$$R_{\text{emp}}[f] := \frac{1}{m} \sum_{i=1}^{m} c(x_i, y_i, f(x_i)).$$

This is a particular case of *M-estimation*.

- The above minimization problem may be an ill-posed problem (example given in the book), since the minimization depends on the function class (overlearning, computational issues etc.).

3.3 A Statistical Perspective – MLE

- Given a particular pattern x, we may want to ask what risk we can expect for it, we need to know the conditional *density* $p(y|x)$.
- An approach is to model the underlying dependence, and maximize the functional form $p(y|x, f(x))$ with respect to x.
- The likelihood of a sample $(x_1, y_1), \ldots, (x_m, y_m)$ given underlying functional dependency f is

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

3.3.1 Maximum Likelihood Estimation (2)

- Minimizing likelihood is equivalent of minimizing (we may drop $p(x_i)$'s since they do not depend on f) the Log-likelihood

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

- (Regression Loss Functions) Minimization of $\mathcal{L}[f]$ and of $R_{\text{emp}}[f]$ coincide if $c(x, y, f(x)) = -\ln p(y|x, f)$. If there is additive noise, i.e. $y_i = f_{\text{true}}(x_i) + \varepsilon_i$, then $c(x, y, f(x)) = -\ln p_\varepsilon(y - f(x))$.

3.3.1 Maximum Likelihood Estimation (3)

- (Classification Loss Functions) Minimization of $\mathcal{L}[f]$ and of $R_{\text{emp}}[f]$ coincide (for binary classification) if

$$c(x, y, f(x)) = \frac{1+y}{2} \ln P(y = 1|f(x)) + \frac{1-y}{2} \ln P(y = -1|f(x)).$$

- Needs modeling of $P(y = 1|f(x)$. A possibility is (logistic) link functions, see the book for details.

3.3.2 Efficiency

- Choice of the loss function is not (of course) arbitrary, and we need tools for comparing estimators.
- An estimator $\hat{\theta}(\mathcal{Y})$ of the parameters θ is simply a function of $\mathcal{Y}$ (independent of θ). An estimator $\hat{\theta}$ of the parameters θ in $P(y|\theta$ is unbiased if $\mathrm{E}\,\hat{\theta} = \theta$. Unbiased estimators are not necessarily "better" than biased.
- A way to compare estimators is their variance. The Fisher Information is given by $I_{ij} := \mathrm{E}[\partial_{\theta_i} \ln p(\mathcal{Y}|\theta) \cdot \partial_{\theta_j} \ln p(\mathcal{Y}|\theta)]$. Denote the covariance of $\hat{\theta}(\mathcal{Y})$ by B. Then any unbiased estimator $\hat{\theta}(\mathcal{Y})$ satisfies (Cramer-Rao) $\det(IB) \geq 1$.

3.3.2 Efficiency (2)

- The efficiency of an estimator is defined by $e := 1/\det(IB)$.
- The closer e is to 1, the lower the variance. For certain loss functions, the efficiency of corresponding estimators can be calculated (see theorem in the book).
- Theorem: The ML estimator is *asymptotically* efficient.

3.4 Robust Estimators

- The idea of robustness is to find estimators that are not sensitive to "bad" observations (outliers). A *median* is an example of a robust location estimator.
- The basic idea in constructing the robust loss functions is to start from ML framework and modify it such that the influence of an individual pattern is limited.
- These ideas lead to Robust Loss Functions, which are studied (in some special cases) in section 3.4.1 and 3.4.2. Finally, some considerations of using these ideas to construct adaptive loss functions, (i.e. the noise model is slightly changing over time).

3.4 Robust Estimators

- The idea of robustness is to find estimators that are not sensitive to "bad" observations (outliers). A *median* is an example of a robust location estimator.
- The basic idea in constructing the robust loss functions is to start from ML framework and modify it such that the influence of an individual pattern is limited.
- The above minimization problem may be an ill-posed problem (example given in the book), since the minimization depends on the function class (overlearning, computational issues etc.).

Problem 1

- (3.1) The soft margin loss function c_{soft} and the logistic loss c_{logist} are asymptotically almost the same; show that

$$\lim_{f \to \infty} \left(c_{\text{soft}}(x, 1, f) - c_{\text{logist}}(x, 1, f)\right) = 1$$

$$\lim_{f \to -\infty} \left(c_{\text{soft}}(x, 1, f) - c_{\text{logist}}(x, 1, f)\right) = 0$$

Problems 2-3

- (3.3) Assume 8 people want to gather for a meeting; 5 of them live in Stuttgart and 3 in Munich. Where should they meet if (a) they want the total distance traveled by all people to be minimal, (b) they want the average distance traveled per person to be minimal, or (c) they want the average squared distance to be minimal? What happens to the meeting points if one of the 3 people moves from Munich to Sydney?
- (3.13) Prove that the least mean square estimator is unbiased for arbitrary symmetric distributions. Can you extend the result to arbitrary symmetric losses.