Note For Data Analysis

Kai Wei

March 2019

1 Introduction

This document is the note for *Data Analysis in High Energy Physics*. Key concepts in statistic and analysis methods in HEP are summarized here. In Chapter 1. fundamental concepts are listed, different distributions are described and two major inferences, frequentist and Bayesian, are compared. In Chapter 2 and 4, methods for parameter and interval estimation are introduced. In high energy physics, hypothesis testing is critical in the interfence of physical conclusions, which contained in Chapter 3. (**FIXME:** following chapters will be implemented later)

2 Fundamental Concepts

Fundamental Concepts introduces the basics of statistical data analyses, such as probability density functions and their properties, theoretical distributions (Gaussian, Poisson and many others) and concepts of probability (frequentist and Bayesian reasoning).

2.1 Probability Density Function

Expectation Value: **Moments**: **Variance**: **Skew and Kurtosis**: **Covariance and Correlation**: **Marginalisations and Projection**: **Other properties**: **Associated Functions**:

This chapter introduces various techniques of event classification. *Events* are classified into either *data* or *background*.

6.1 Introduction to Multivariate Classification

multivariate classification methods or *algorithms* (MVA) use the multi-dimensional observable space rather than each observable separately.

Key Idea: In general, multivariate techniques turn the *feature vector* $\boldsymbol{x}$ into one single variable γ. Multivariate selection algorithm as a mapping function, $\Re^D \to \Re : \gamma = \gamma(\mathbf{x} = x_1, \ldots, x_D)$. Each $\gamma = c$ corresponds to a hypersurface in observable space. Classification will be made based on this single variable.

Two approaches of realization:

a). To pproximate the true pdf ($p(x|S)$ and $p(x|B$) using Monte Carlo-simulated events.

b). To construct a suitable variable which can be used as a multivariate classifier.

Analyses can loosely be categorized as:

• precision measurements: These require high purity p, that is a large fraction of signal events in the selected sample. This is achieved by keeping type I errors small;

• trigger selections: These need high efficiency $\epsilon = 1 - \beta$ which is achieved by keeping type II errors small;

• cross-section measurements: These are optimized by maximizing the signal significance which is often approximated by $\nu_s/\sqrt{\nu_s + \nu_b}$(or equivalently by maximizing the quantity $\sqrt{\epsilon \cdot p}$);

• searches for new particles: These typically have $\nu_s \ll \nu_b$ and are optimized by maximizing $\nu_s/\sqrt{\nu_b}$.

decision boundary: the boundary of the critical region C.

Receiver-Operating-Characteristic Curve: type I and type II can be made arbitrarily small individually. The best working point (the cut $\gamma = c$) has to be found for each analysis purpose.

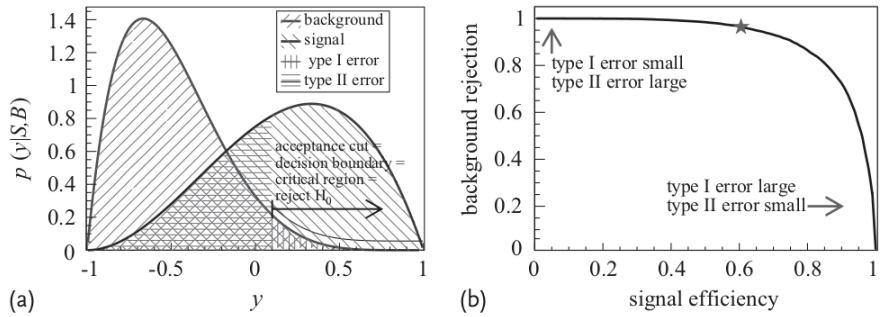

Figure 5.2 (a) The distributions $p(y|S)$, $p(y|B)$ of an MVA variable y for signal and background events. The classification is based on a cut on the MVA variable y which in this example is chosen to be at $y = 0.1$. (b) The ROC curve, showing the background rejection as a function of the signal efficiency achieved by varying the cut on the MVA output variable. The working point according to the cut value of 0.1 is indicated by a star.

Figure 1: ROC curve

Receiver-Operating-Characteristic (ROC) curve is used to characterize the performance of a classification algorithm, which shows the relation between the signal efficiency and the background rejection. See 1

Neyman-Pearson Lemma: A classification algorithm based on the likelihood ratio

$$\gamma(x) = \frac{p(x|S)}{p(x|B)} \tag{1}$$

as test statistic results in the critical region with the largest signal efficiency $1 - \beta$. Neyman-Pearson lemma guarantees that *Most Powerful Test* (MPT) exists.

Supervised Machine Leaning: automated determination of the decision boundary according to a chosen algorithm.

Bias-Variance Trade-Off: Defining the necessary flexibility of the model and fixing the model configuration parameters is an optimization procedure which help to find the balance point between bias and variance.

Cross-Validation: when training events are limited for training and validation, one can apply cross-validation. Training sample T is split into K independent subsets T_k, and then K classifiers using the same configuration parameters are trained on each of the sets $T \backslash T_k$. Then T_k is used to validation and optimize the configuration.

6.2 Multivariate Classification Techniques

• ***Likelihood (Naive Bayes Classifier)***:

Assuming the absence of correlations between the observables, multidimensional pdfs is then the production of the one-dimensional pdfs $p_{s(b),k}(x_k^{(i)})$. The

resulting classifier is called the *likelihood classifier*, $L_{s(b)}$.

$$L^{(i)}_{s(b)} = \prod_{k=1}^{D} p_{s(b),k}(x^{(i)}_k) \tag{2}$$

As the classifier one uses the likelihood ratio $\gamma^{(i)}_L$ for event i, which is defined by

$$\gamma^{(i)}_L = \frac{L^{(i)}_s}{L^{(i)}_s + L^{(i)}_b} \tag{3}$$

With larger dimension, the fluctuation on event density will be amplified by large number of dimension, which is called the *curse of dimensionality*.

• ***k-Nearest Neighbour and Multi-dimensional Likelihood***: *k-Nearest Neighbour* (kNN) algorithm provides an estimate for a multi-dimensional likelihood where the distribution density is approximated by the number of signal and background events of the training sample, $k_s(\mathbf{x})$ and $k_b(\mathbf{x})$. The likelihood ratio can be written as:

$$\frac{p(\mathbf{x}|S)}{p(\mathbf{x}|B)} \propto \frac{p(S|\mathbf{x})}{p(B|\mathbf{x})} \simeq \frac{k_s(\mathbf{x})}{k_b(\mathbf{x})} \tag{4}$$

This algorithm adopts the size of the region over which the pdf is averaged to the available training data. But we still need to specify the *metric* that defines the distance. A good choice is *Mahalanobis distance*.

$$R_{rescaled} = [\sum_{k,l=1}^{D} (x_k - y_k)(\mathbf{V}^{-1})_{k,l}(x_l - y_l)]^{\frac{1}{2}} \tag{5}$$

while $\mathbf{V}^{-1}$ is a full inverse covariance matrix.

7 Unfolding

8 Constrained Fit

9 How to deal with Systematic Uncertainty

10 Theory Uncertainty

11 Statistical Methods in HEP

12 Analysis Walk-Through

13 Application in astronomy