Advanced Multivariate Methods in High Energy Physics

Mikael Mieskolainen

Helsinki Institute of Physics

ECT* Trento, 29/02/2012

Motivation for Multivariate methods?

If we have *a rare* or *experimentally* difficult scattering process or decay, multivariate methods should be near optimal way to do statistical inference (if we have a good MC model)

- Reductive cuts do not fully utilize correlations between variables

Some applications in HEP

- Discriminating physics signal and physics background
- Multi-class classification (e.g. single/double/non-diffraction)
- Track reconstruction/fitting
- Particle identification
- Energy regression (CMS-2011)
- Triggering!

Methods are usually always *supervised* and trained using a known sample set, *unsupervised* techniques also exists...

Regression or Classification?

Let us have $d \in \mathbb{N}$, real-valued physically measured variables

Classification is a mapping

$$g_C : \mathbb{R}^d \to \mathcal{C}, \tag{1}$$

where $\mathcal{C}$ is a *discrete* set of class labels, e.g.
$\mathcal{C} = \{1 := \textit{signal}, 2 := \textit{background}\}$

Regression is a mapping

$$g_R : \mathbb{R}^d \to \mathbb{R}^n, \tag{2}$$

where $n \in \mathbb{N}$, usually $n = 1$

Multidimensional likelihood

So we have a real-valued d-dimensional *continuous random vector* $\mathbf{X} \in \mathbb{R}^d$. Especially, we assume that there is a *total likelihood* function $f_{\mathbf{X}} : \mathbb{R}^d \to [0, \infty)$ such that there exists probability

$$P(\mathbf{x} \in A) = \int_A f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}, \tag{3}$$

where $A \subset \mathbb{R}^d$ is an interesting region of sample space. Integral of $f_{\mathbf{X}}$ over total sample space Ω should be one.

Probabilities

Formally, we get a posteriori classification probabilities of a measurement vector $\mathbf{x}$ from the Bayes' rule using class likelihoods f_j and priors P_j

$$P(C = j|\mathbf{X} = \mathbf{x}) = \frac{f_{\mathbf{X}}(\mathbf{x}|C = j)P(C = j)}{f_{\mathbf{X}}(\mathbf{x})} = \frac{f_j(\mathbf{x})P_j}{\sum_{j'=1}^{|\mathcal{C}|} f_{j'}(\mathbf{x})P_{j'}}, \quad (4)$$

where $j = 1, \ldots, |\mathcal{C}|$. It the task of a multivariate algorithm to estimate these probabilities.

Bayesian prior probabilities P_j can have a major role!

Supervised classification

In practise, learning supervised classifier is based on a training set $\mathcal{T}$ of continuous measurement vector - discrete class label pairs:

$$\mathcal{T} = \{(\mathbf{x}_i \in \mathbb{R}^d, c_i \in \mathcal{C})\}$$

- Usually training vectors are generated for different classes using MC (biggest drawback)

Why probabilistic methods?

Because class likelihood functions are often inherently overlapping even in d-dimensional space!

Reason for this can be physics process based, or experimental limitations like detector resolution, limited geometry or electronic noise. Effects can be linear (convolutive), non-linear, additive etc.

That is why multivariate algorithms estimating also *a posteriori probabilities* of classifications are important

Standard methods like Neural Networks (NN) and Boosted Decision Trees (BDT) do *not* really give any probabilities

Likelihoods and posteriors

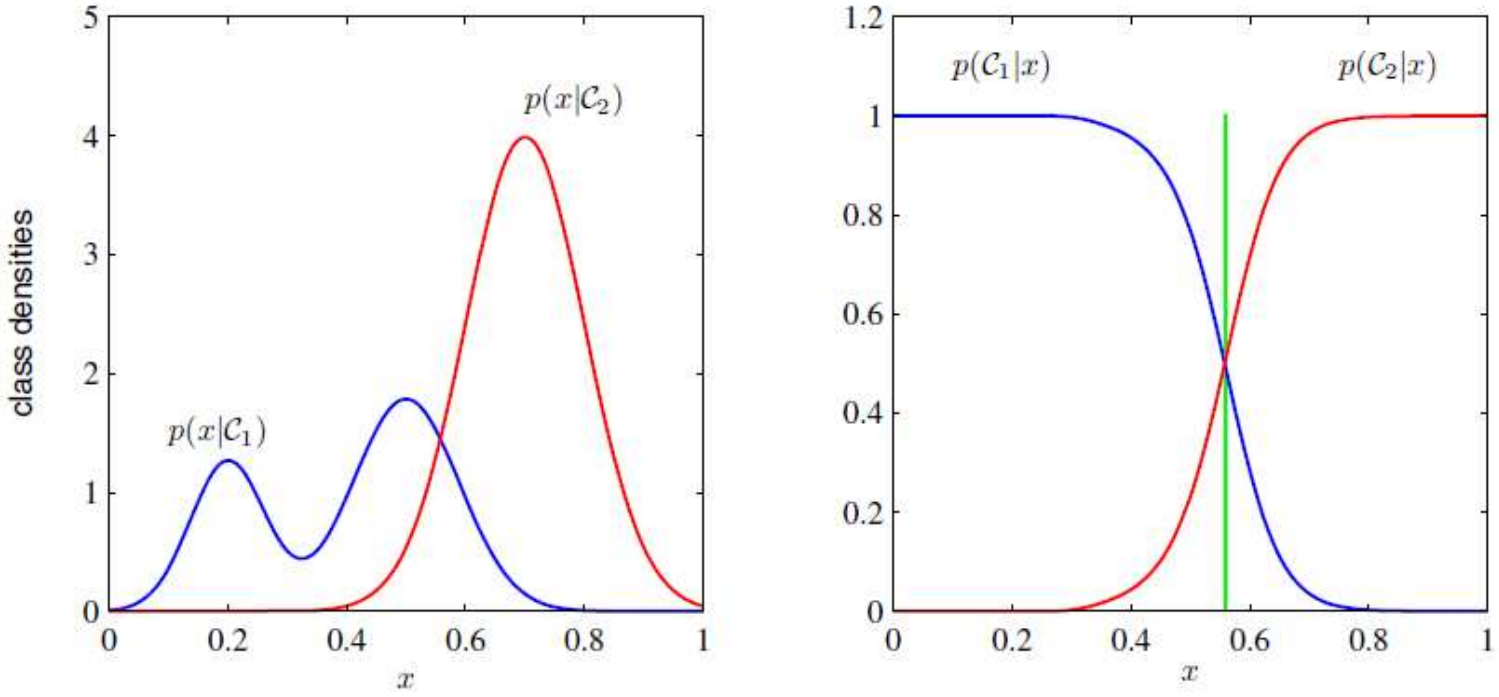

Figure: Class likelihood functions on the left, a posteriori probabilities on the right. [C.M. Bishop, Pattern Recognition]

Hard classifiers

So most practical classifiers do *not* estimate posterior probabilities but only define *decision boundaries* in space $\mathbb{R}^d$. These decision boundaries can be affine hyperplanes or in general, nonlinear surfaces.

One probabilistic classifier

A posteriori probabilities of an event vector $\mathbf{x}$ can be estimated with ℓ_1-norm regularized Multinomial Logistic Regression (MLR-ℓ_1)

$$P(C = j|\mathbf{x}) = \frac{\exp(\langle \mathbf{w}_j, \mathbf{x} \rangle)}{\sum_{j'=1}^{|\mathcal{C}|} \exp(\langle \mathbf{w}_{j'}, \mathbf{x} \rangle)}, \quad j = 1, \ldots, |\mathcal{C}|,$$

where weight vectors $\mathbf{w}_j$ are trained with the training set $\mathcal{T}$ using mathematical convex optimization

- Regularization of the cost functional in training phase with ℓ_1-norm induces *sparsity* into $\mathbf{w}_j \Rightarrow$ algorithmic variable (detector) selection

Example of regularization: Diffraction classification

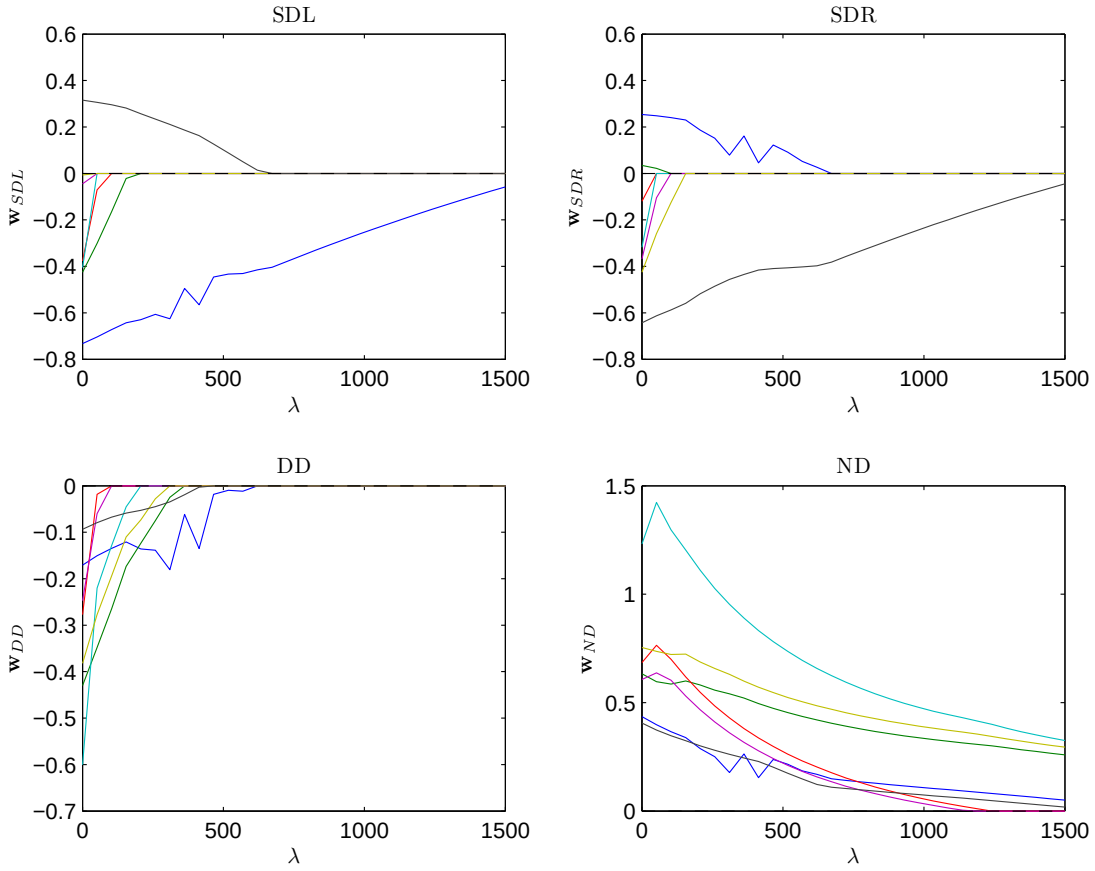

Figure: Regularization λ paths using MLR-ℓ_1, variables of $\mathbf{x}$ are calorimeter energies, PYTHIA6, CDF. On y-axis; coefficients of $\mathbf{w}_j$ in order: $w_1 :=$ blue ($\eta = -3.6$), green, red, purple, magenta, brown, $w_7 :=$ black ($\eta = 3.6$)

Summary

- We should head toward multivariate methods with probabilistic output
- More transparent, *non* black-box methods should be favoured
- Pre-cuts, variable normalization, uncertainties and biases of methods are non-trivial things ...

Thank you!