Machine learning techniques in LHCb

Patrick Owen
MLHEP summer school
28/08/15
Jargon transformation

- Overfitting —> Overtraining
- Feature —> Variable
- Model —> classifier
- Bias/correlation to POI —> feature selection (?)
- Model ensembling —> ‘blending’
- k-Fold cross-validation —> k-Fold cross-validation (?)
The LHCb experiment
The LHCb experiment

Overall view of the LHC experiments.

- **VELO**: Primary vertices, impact parameter
- **RICHES**: K, pi particle ID
- **E/HCAL**: Trigger, p, e, gamma PID
- **MUONS**: Trigger and PID
- **TRACKER**: P of charged particles
- **Magnet**

Source: LHCb-PUB-2012-017
Physics beyond the SM

• There are many problems with our theoretical description of the universe, the Standard Model (SM).

• Only 5% of matter accounted for by SM.

• SM CP violation \(\sim 10\) orders of magnitude too small to account for observed matter/anti-matter asymmetry.

• Nothing to do with gravity.

• + some fine tuning problems.

• The hope is that we find some new particles at LHC to fix some of this.
The goal of LHCb

- Core physics programme: Indirect searches for new physics in heavy flavour decays.

- Example: $B_s^0 \rightarrow \mu^+ \mu^-$

- Measuring properties of B decays can tell us about NP way beyond accelerator energy.
Hadronic environment

• Many precise B physics results have been made at $e^+e^-$ colliders.
  
• Very clean environment to work in, typically only 2B mesons produced in the detector.

• Roughly 1000 times more B-mesons produced so far at LHCb interaction point compared to B-factories

• However, hadronic environment much more difficult to work with.
Outline

• Lower level information:
  • Particle ID
  • Tracking
  • Flavour tagging

• Physics analyses (most use MVA):
  • Canonical example: $B^0_s \rightarrow \mu^+ \mu^-$
  • Bias to observables of interest: Dark boson search
  • Badly modelled variables: $B^0 \rightarrow K^{*0} \mu^+ \mu^-$
  • Choice of classifier: $\tau^+ \rightarrow \mu^+ \mu^- \mu^+$
Particle ID at LHCb

• Several sub-detectors provide particle ID information.
RICH detectors

- Particles that travel faster than the speed of light will emit a cone of light, known as Cherenkov radiation.

  \[ \theta_C = \cos^{-1}\left(\frac{1}{\beta n}\right) \]

- Allows LHCb to distinguish between pions and kaons.

- Criteria used in analysis combines likelihood information from each sub detector (combined DLLs).
Particle ID in action

- Particle ID crucial for most LHCb analyses.

From RICH performance paper, arXiv:1211.6759

- Most commonly used criteria is the combined DLL information.
Limitations with combined DLL

- Not all information can be represented as a likelihood.
- Binary, or discrete variables, e.g. detector acceptance flags.
- Combining likelihoods not always mathematically defined (sometimes need to re-scale input).

- Instead, use MVA.
- Train on simulation (all tracks).
- Combine tracking information with PID information.
Performance compared with DLLs

- Various models are tested, and the Multi-layered perceptron is most performant.

- MVA looks to be much more performant than CombDLL.

- However, this problem is not always what analysts want to solve.

- They often really care about K-pi separation, in that case often the CombDLLs have similar performance, despite plot on right.

- MVA PID better at getting rid of combinatorial background.
Tracking in LHCb

- Measure direction of tracks before and after dipole magnet - measure momentum.

- Tracks are formed by adding hits which are consistent with a parabola.

- If noise hits are added, or hits from other particle, it is called a ‘ghost’.
Ghost probability tool

- Artificial neural network.
- Trained on simulation.
- Use various properties of the track (# hits in sub-stations, quality etc ..).

Figure 1: Distribution of input variables to the classifier for good tracks (blue) and ghosts (red) in the training sample.

Figure 3: Distribution of ghost probability output values for good tracks and ghosts on long tracks.
Flavour tagging

• Neutral B mesons oscillate into their anti-partners, and vice versa.

• Flavour tagging is the ability to tell the flavour of a neutral B meson at the time of production.

• Crucial for probing NP inside these box diagrams.
How it works

Same side (SS): Fragmentation of b-quarks leads to correlation between spectator quark and nearby particles

Opposite side (OS): Exploit the fact that there are always two b-quarks produced with opposite flavour.
Same-side tagging

• Tagging is a classification problem:
  • Distinguish between particles originating from fragmentation (signal), and those from elsewhere (background).

• Select several good quality tracks, with kaon PID from the RICH detectors.

• Choose tagging particle based on how close it is to the Bs phase-space (pseudo-rapidity difference, phi, BsK mass etc).
Neutral Net tagger

- Initially cut based selection used to choose tag particle [LHCb-CONF-2012-033].

- Now use Neutral Net, combining information from the entire event (track multiplicity particularly useful).

- Improvement in tagging power equivalent to having 40% more data!

- Can calculate mis-tag probability on event-by-event basis.

Btw, these numbers never make any sense.
Opposite side taggers

- To exploit opposite side, need to think about common B decays.
  
  - $B \rightarrow DX \ 85\%$
  
  - $B \rightarrow \ell \nu X \ 25\%$

- Find flavour of lepton/charm to get flavour of other B in event.

- Diluted by opposite side oscillations.
  
  - 30 (100)% $B^0(B_s^0)$ oscillate before they decay
OS vertex tagger

• A more inclusive way to is to calculate charge of a detached vertex on other side.

• Best detached vertex chosen by NN classification.

• Tracks added to this candidate, NN response updated.

• Calculate vertex charge, mark as untagged if too low.

\[ Q_{vtx} = \frac{\sum_i Q_ip_{Ti}^k}{\sum_i p_{Ti}^k} \]
\[ B^0_s \rightarrow \mu^+ \mu^- \]

Ultra-rare decay confirmed in LHC

By Melissa Hogenboom
Science reporter, BBC News

◎ 24 July 2013 | Science & Environment
Huge background, small signal

- Let's look at the expected rate for $B^0_s \rightarrow \mu^+ \mu^-$

$$\mathcal{B}(B^0_s \rightarrow \mu^+ \mu^-) \sim 3 \times 10^{-9}$$

- Background originates from two separate $B$ decays, both decaying

$$\mathcal{B}(\bar{b} \rightarrow \mu^+ X) \times \mathcal{B}(b \rightarrow \mu^- X) \sim 1 \times 10^{-2}$$

- Before any selection, the background is seven orders of magnitude bigger than the signal.
Training samples

- First job, decide on samples to train on, which represent signal and background.

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Real data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easily test for bias</td>
<td>Guaranteed to be realistic</td>
</tr>
<tr>
<td>No worry about overtraining</td>
<td>Typically bigger than</td>
</tr>
<tr>
<td></td>
<td>simulation samples</td>
</tr>
</tbody>
</table>

- For $B_s^0 \rightarrow \mu^+ \mu^-$, understand background well enough to use simulation for both.
Calibration

- Calibrate response of MVA on $B \rightarrow hh$, same topology.
- In order for this to work, omit particle ID from list of variables.
- For background, can use data outside the signal region to check response.
Signal discrimination: BDT

Goal is to differentiate signal events from combinatorial background.

BDT training, choice of variable and BDT parameters optimization based on MC signal and background (new sample equivalent to 7 fb−1).

12 variables used (previously 9) based on kinematic and topological information chosen to avoid correlation with invariant mass.

Bs2MuMu @ LHCb

Desirable variable qualities:
- Have physical reason to distinguish
- Well reproduced in simulation
- Not correlated with dimuon mass
Variables

BDT variables

B candidate:
- proper time
- IP
- $p_T$
- isolation
- Angle between the $B$ momentum and $P$ thrust
- Angle between $\mu^+ \mu^-$ direction in the $B$ rest frame and $P$ thrust in the $B$ rest frame

$P$ thrust is the sum of momenta of all tracks consistent with originating from the decay of the other $b$ hadron

Muons:
- min IP significance
- distance of closest approach
- isolation
- polarization angle
- $|\eta(\mu_1) - \eta(\mu_2)|$
- $|\phi(\mu_1) - \phi(\mu_2)|$

Diagram showing angles and distances related to the $B_s^0 \rightarrow \mu^+ \mu^-$ decay, with scatter plots illustrating the distributions of various variables for data and simulation.
Variables

**Pseudo-Tau Variables**

- Proper time
- **IP**
- **p**
- Isolation
- Angle between the **B** momentum and **P**
- Angle between µ^+ direction in the **B** rest frame and **P** in the **B** rest frame

**P** is the thrust of all tracks consistent with originating from the decay of the other b hadron.

**Muons**

- Minimum Ip significance
- Distance of closest approach
- Isolation
- Polarization angle
- |η(µ_1) - η(µ_2)|
- |φ(µ_1) - φ(µ_2)|

**DOCA (%)**

- **Beam line**

**Graphs**

1. LHCb:
   - Data sidebands 3 fb^-1
   - Data exclusive B→hh' 3 fb^-1
   - Simulation bkg
   - Simulation signal

2. LHCb:
   - Data sidebands 3 fb^-1
   - Data exclusive B→hh' 3 fb^-1
   - Simulation bkg
   - Simulation signal

3. LHCb:
   - Data sidebands 3 fb^-1
   - Data exclusive B→hh' 3 fb^-1
   - Simulation bkg
   - Simulation signal

4. LHCb:
   - Data sidebands 3 fb^-1
   - Data exclusive B→hh' 3 fb^-1
   - Simulation bkg
   - Simulation signal
Isolation

- Form cone around B direction, add $p_T$ of all other tracks around this cone.

\[
I = \frac{p_T(B)}{p_T(B) + \sum_{i}^{\text{tracks}} p_T^i}
\]

- Now we see some disagreement between calibration and training samples - is this OK?

- Depends on situation, here removing background is much more important than anything else.
Multivariate isolation

- Many LHCb publications now use separate MVA to isolate signal from nearby particles.

- Most important variable is the vertex quality difference when a track is added to signal vertex.
MVA application

- Simplest option is to place a cut on BDT response.
- Most analyses in LHCb do this.
- Typically optimise requirement by maximising a figure of merit.

\[
\frac{S}{\sqrt{S + B}} \cdot \frac{\epsilon}{N_{\sigma}/2 + \sqrt{B}}
\]

\[
\mathcal{L}(\text{sig} + \text{bkg}) / \mathcal{L}(\text{bkg})
\]
Binned response

• Will always get more sensitivity by binning the response instead of cutting on it.

• All events above BDT cut are treated equally, but this is of course not the case.

• There is no downside to doing this method, apart from the additional complexity.
Final performance

• Difficult to directly compare with cut-based approach (dropped it long ago), but clear that MVA made discovery possible.

• LHCb+CMS published observation of the decay in Nature earlier this year.
Choice of classifier

- Within LHCb, most common choice of classifier is BDT or Neutral Net.

- For $B_s^0 \rightarrow \mu^+ \mu^-$, there were many that were tried, performance based on separation power and bias to di-muon mass.

- Hot competition for the last round led to a very thorough optimisation, good improvement compared to previous papers.

- Proper tuning of configuration and experience tends to be more important than the actual choice.
Blending method

- Interesting approach used by $\tau^+ \rightarrow \mu^+\mu^-\mu^+$ analysis
- Train 8 different classifiers with different configurations and combine responses into one BDT classifier.
  - 1 BDT, 2 Fisher discriminants, 4 neural networks, 1 one function-discriminant analysis, 1 linear discriminant.
- Improves sensitivity by 6% compared to using the best single classifier.

Badly modelled variables

• Consider a slightly different decay, $B^0 \rightarrow K^{*0} \mu^+ \mu^-$

• Not as rare as $B^0_s \rightarrow \mu^+ \mu^-$, (no helicity suppression).

• Get an extra pion and kaon in the final state.

• Very important to use PID information in selection to remove background
Using Particle ID in an MVA

Particle identification criteria dependence strongly on momentum - perfect example of a variable that is useful in an MVA

Unfortunately, it is not perfectly well modelled in the simulation. Cannot correct by simple 1D re-weighting - correlated to other observables.
Particle ID

- Use calibration samples in data to generate PID distributions as a function of momentum, eta, and track multiplicity.
- Lose correlations between different PID observables.
Training on data

- Unfortunately unlike simulation, data is not purely signal.

- We use s-Plot technique to subtract background from signal region.

- Negatively weight the events in background rich regions to get purely.
Keeping data unbiased

• Obviously can’t train (or test) on data if you plan on using it later.

• Common solution in LHCb is to split data sample into subsets.

• Train/test on n-1 subsets, apply to final subset.

• Technique first used in LHCb for the $K^0_s \rightarrow \mu^+ \mu^-$ analysis.

LHCb-PAPER-2012-023, arXiv:1209.4029
Correlation with mass

• Nearly always want to fit a mass distribution after MVA selection.

• Need to keep correlation between MVA and mass small, and smooth.

• Typically only an issue with two-body decays, as its easy to calculate the mass of the parent.
Slightly more complicated example

• LHCb good place to search for light, exotic particles, present in many extensions to the SM.

• Unknown lifetime and mass - want to train a selection which is unbiased in both of these.

• Problem: almost every useful variable is going to be correlated to at least of one these.
uBoost

- uBoost, is a method of boosting which preserves a uniform efficiency in variables of choice.

- Multiply boosting weights by the non-uniformity.

- Approach improved recently, see: A. Rogozhnikov et al, arXiv:1410.4140

- Not only mis-classified events are boosted but also events which have low efficiency.
uBoost advantages

• To avoid bias normally one needs to remove variables

• Exploit maximal amount of information whilst keeping POIs flat in efficiency.

• uBoost BDT has flat efficiency in both mass and lifetime.
Results

- Set limits on decay mode.

- Efficiency as a function of lifetime inversely proportional to limit, but no issue regarding final selection.
things for the future

conservatism

[kuh n-suh-voe-tiz-uh m]

Examples

noun

1. the disposition to preserve or restore what is established and traditional and to limit change.
Resolution and regression

- Many interesting $B$ decays contain neutrinos in the final state, e.g. $B \rightarrow D \mu \nu$.

- Two methods commonly used: k-factor and ‘neutrino reconstruction’.

\[
\text{Fit the corrected mass:} \\
M_{\text{corr}} = \sqrt{p_{\mu}^2 + M_p^2} + p_{\nu}^2 \\
\text{Determine its uncertainty.} \\
\text{Reject candidates if:} \\
M_{\text{corr}} > 100 \text{ MeV}/c^2 \\
\text{This does not utilise any information rest of event - could a regression algorithm improve things further?}
\]
Image recognition software

- Most of the ML tools we use are fairly old now.

- Can we make use of image recognition software? For isolation? For flavour tagging?
Conclusion and outlook

• LHCb is an LHC experiment dedicated to studying charm and beauty particles.

• Machine learning is used in many places in LHCb.

  • It's rare to see an analysis without an MVA.

• Mostly use ML for high level information (momentum, impact parameter ..).

• We are still using old MVA techniques, but are slowly moving towards more modern tools and techniques.