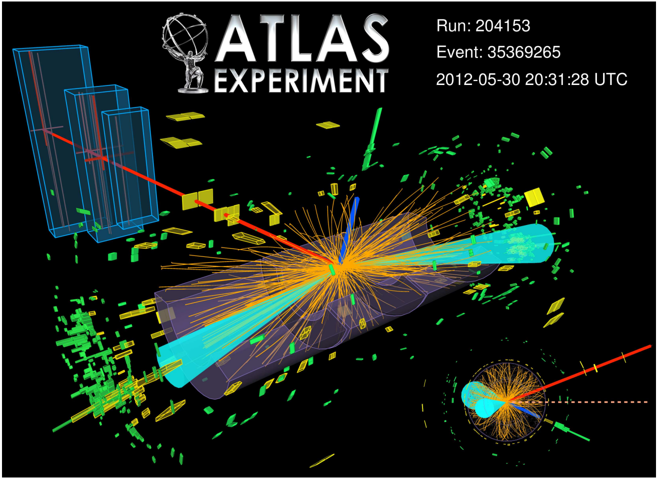
Usage of SVM for a Triggering Mechanism for Higgs Boson Detection

Klemen Kenda, Dunja Mladenić

Artificial Intelligence Laboratory, JSI

October 9th, 2017

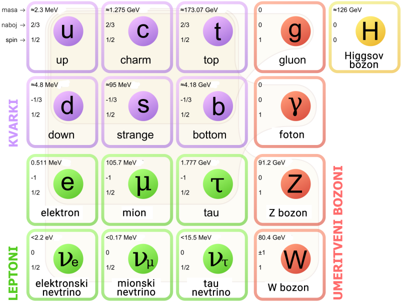
Motivation


ATLAS EXPERIMENT
Run: 204153
Event: 35369265
2012-05-30 20:31:28 UTC

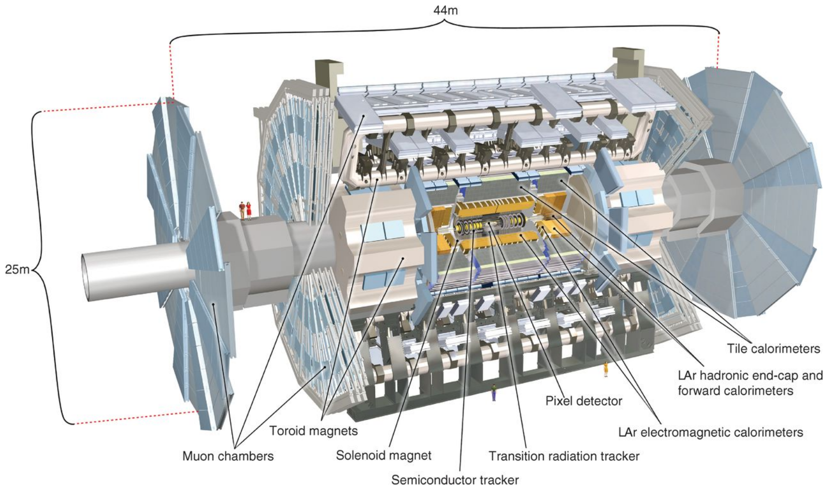
Agenda

Higgs Boson Quest

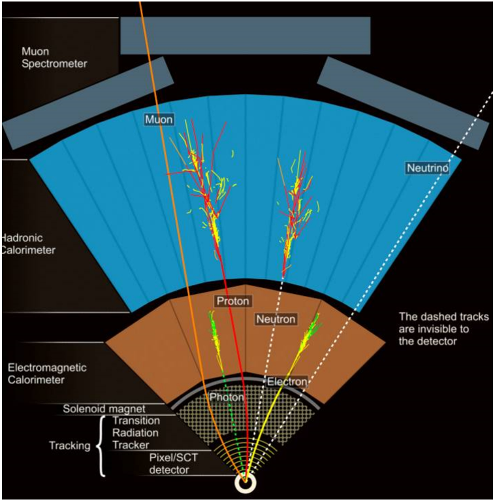
- 1964 - existance predicted (Higgs, Englert, Brout)
- 2012 - ATLAS and CMS
- 2013 - Nobel Prize for Physics (Higgs, Englert)



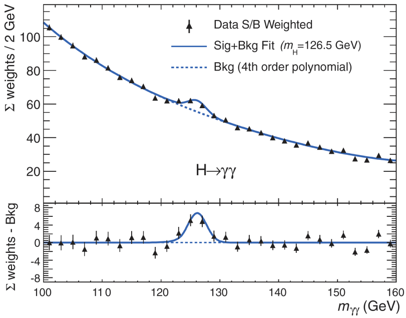
ATLAS Detector 1/2

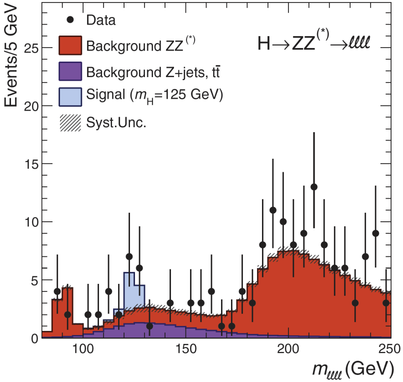


ATLAS 2/2 Detector

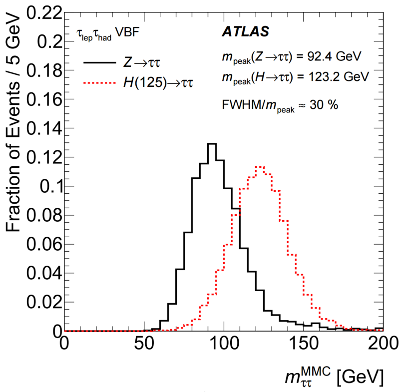


$H \to \gamma\gamma$ in $H \to Z^0 Z^0 \to \ell\ell\ell\ell$





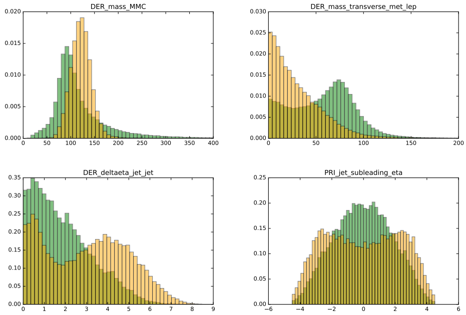
$H \to \tau^+\tau^-$



- $H \to \gamma\gamma$
- incomplete final state
- special decay topology: first τ in e^- or μ^- and 2 ν, second τ into hadrons and 1 ν
- Public challenge *HiggsML*

Data

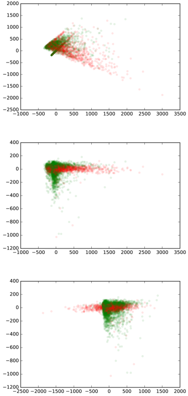
- ATLAS detector simulator
- 250.000 events (1/3 signal)
- 30 features (primary and derived)
- id
- weight
- exploratory data analysis



Histograms of typical features.

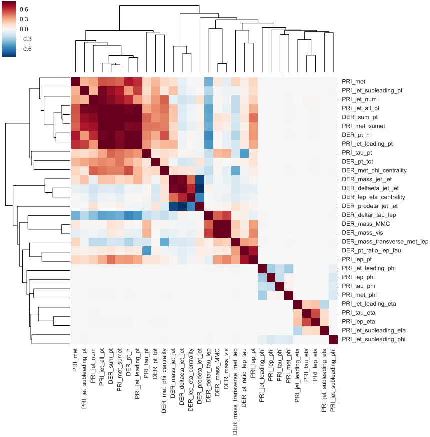
Exploratory Data Analysis

PCA



On x-axis there is the first, and
on y-axis the coordinate of second component.

Feature correlations



Problem Definition

Training Set:

$\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}, w^{(1)}), \ldots, (\mathbf{x}^{(n)}, y^{(n)}, w^{(n)})\}$

Legend

$\mathbf{x}^{(i)} \in \mathbb{R}^d$;feature vector, d - number of features
$y^{(i)} \in \{\mathrm{b}, \mathrm{s}\}$; label, b - background, s - signal
$w^{(i)} \in \mathbb{R}^+$; weight - probability for an event

The Problem

We are looking for classifion function $g : \mathbb{R}^d \rightarrow \{\mathrm{b}, \mathrm{s}\}$, that will yield best classification $x^{(i)}$ on the training (validation) set according to the selected metrics.

Metrics and evaluation

- precision $PPV = \frac{TP}{TP+FP}$
- recall $TPR = \frac{TP}{TP+FN}$
- F_1 score
  $F_1 = 2 \cdot \frac{PPV \cdot TPR}{PPV+TPR}$
- AMS_2 metrics

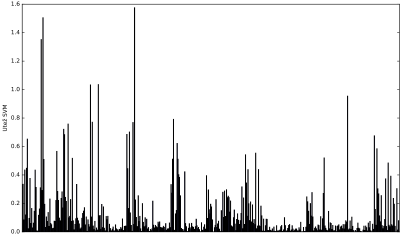
- splitting the training set
- test set
- overfitting

$$\text{AMS} = \sqrt{2\left((s + b + b_{reg}) \ln\left(1 + \frac{s}{b + b_{reg}}\right) - s\right)}$$

$$s = \sum_{i \in \mathcal{S} \cap \hat{\mathcal{G}}} w^{(i)} \quad b = \sum_{i \in \mathcal{B} \cap \mathcal{G}} w^{(i)}$$

Feature generation 1/2

- Non-linearity (boosting and deep learning give best results)
- One or two- features transformation,
- Functions: x^2, x^3, e^x, $\sqrt{x}$, $\log(x)$, missing values
- Cluster id as a new feature
- Feature selection



SVM weights per feature with combinations of tranfsormations for 2 features

Feature Generation 2/2

Id	Description
(1)	Original set.
(2)	Added missing values.
(3)	Filtered missing values and all e^x.
(4)	Filtered missing values and all x^2.
(5)	Filtered missing values, x^2 and all x^3.
(6)	Filtered missing values, x^2, x^3, e^x and all $\sqrt{x}$.
(7)	Filtered missing values, x^2, x^3, e^x, $\sqrt{x}$ and all $\log(x)$.
(8)	Selection of best features by one transformed variable: missing values, x^2, e^x, $\sqrt{x}$.
(9)	Unfiltered set by one variable transformed values.
(10)	Unfiltered set $x_i x_j$ (435 features).
(11)	Feature set from HiggsML Challenge winner - Tim Salimans.
(12)	Unfiltered set $x_i^2 + y_i^2$.
(13)	Unfiltered set $e^{x_i^2 + y_i^2}$.
(14)	Unfiltered set $\sqrt{x_i^2 + y_i^2}$.
(15)	Unfiltered set $(1 + x_i x_j)^2$.
(16)	Filtered set of one and two-variable transformations.
(17)	(8) with added cluster id.

Table: Used feature sets for SVM.

Linear SVM

Kernel (features)	true positive	true negative	false positive	false negative	precision	recall	accuracy	F_1	AMS_2
LIN (1)	18,9%	56,0%	9,5%	15,6%	0,665	0,548	0,749	0,600	1,999
LIN (3)	22,5%	58,0%	7,6%	11,9%	0,748	0,655	0,805	0,698	2,526
LIN (4)	22,5%	58,0%	7,6%	11,9%	0,748	0,654	0,805	0,698	2,528
LIN (5)	22,6%	57,6%	7,9%	11,8%	0,740	0,657	0,802	0,696	2,478
LIN (6)	23,5%	57,4%	8,2%	10,9%	0,743	0,683	0,809	0,712	2,547
LIN (7)	23,8%	56,9%	8,6%	10,7%	0,734	0,690	0,807	0,711	2,516
LIN (8)	23,1%	57,1%	8,5%	11,4%	0,732	0,670	0,802	0,700	2,482
LIN (10)	24,3%	57,2%	8,3%	10,1%	0,744	0,705	0,815	0,724	2,582
LIN (11)	20,1%	56,7%	8,9%	14,3%	0,694	0,584	0,768	0,634	2,201
LIN (12)	24,3%	57,2%	8,3%	10,1%	0,744	0,705	0,815	0,724	2,583
LIN (13)	24,4%	57,2%	8,4%	10,0%	0,744	0,709	0,816	0,726	2,581
LIN (14)	24,3%	57,2%	8,3%	10,1%	0,744	0,705	0,815	0,724	2,583
LIN (15)	24,4%	57,1%	8,4%	10,0%	0,744	0,710	0,816	0,726	2.578
LIN (16)	23,6%	57,3%	8,3%	10.9%	0,740	0,684	0,809	0,711	2,553

Table: Linear SVM results.

Kernels

Kernel (features)	true positive	true negative	false positive	false negative	precision	recall	accuracy	F_1	AMS_2
POLY, $c = 0$ (1)	19,1%	59,7%	5,9%	15,3%	0,764	0,555	0,787	0,643	2,345
POLY, $c = 1$ (1)	24,1%	57,6%	8,0%	10,4%	0,752	0,699	0,817	0,724	2,632
POLY^2, $c = 1$ (1)	23,8%	58,0%	7,5%	10,6%	0,759	0,691	0,818	0,723	2,633
POLY^4, $c = 1$ (1)	24,0%	57,0%	8,5%	10,4%	0,738	0,698	0,811	0,718	2,570
POLY, $c = 1$ (8)	24,5%	57,7%	7,8%	9,9%	0,758	0,711	0,822	0,734	2,704
POLY, $c = 1$ (11)	24,2%	58,0%	7,5%	10,3%	0,763	0,702	0,822	0,731	2,701
POLY, $c = 1$ (16)	24,3%	57,8%	7,8%	10,1%	0,757	0,706	0,821	0,731	2,640
RBF (1)	23,7%	58,1%	7,4%	10,7%	0,761	0,689	0,819	0,724	2,674
RBF (2)	23,8%	58,2%	7,4%	10,6%	0,763	0,691	0,820	0,725	2,688
RBF (8)	24,4%	58,2%	7,3%	10,0%	0,769	0,708	0,826	0,737	2,783
RBF (11)	23,5%	58,3%	7,2%	11,0%	0,766	0,682	0,819	0,721	2,703
RBF (16)	24,1%	58,3%	7,3%	10,3%	0,768	0,700	0,824	0,732	2,747
RBF (17)	24,7%	58,1%	7,5%	9,8%	0,767	0,716	0,827	0,740	2,772
*RBF (8)	24,7%	59,0%	6,5%	9,7%	0,791	0,718	0,837	0,752	2,940
*RBF (16)	24,5%	59,0%	6,6%	10,0%	0,787	0,711	0,834	0,747	2.916

Table: SVM results with different kernels.

Evaluation

Method (features)	true positive	true negative	false positive	false negative	precision	recall	accuracy	F_1	AMS_2
simple window	28,4%	43,2%	22,3%	6,1%	0,560	0,824	0,716	0,667	1,579
log. regression (1)	18,4%	56,4%	9,1%	16,1%	0,668	0,535	0,749	0,594	2,015
SVM-LIN (16)	23,6%	57,3%	8,3%	10.9%	0,740	0,684	0,809	0,711	2,553
bag. SVM-RBF (8)	24,4%	58,7%	6,9%	10,0%	0,780	0,708	0,831	0,743	2,854
GBC (8)	24,2%	59,0%	6,5%	10,2%	0,787	0,703	0,832	0,742	2,856
SVM-RBF (8)	24,7%	59,0%	6,5%	9,7%	0,791	0,718	0,837	0,752	2,940
XGBoost (1)	27,8%	51,6%	14,0%	6,7%	0,665	0,806	0,793	0,729	3,735

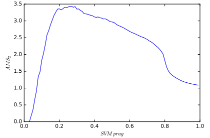
Table: Comparision of results, based on different methods.

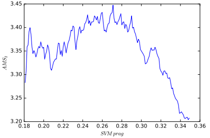
Possible Improvements

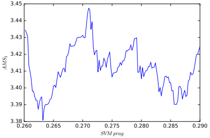
- Possible improvements:
  - Additional threshold optimization - nearer to AMS metrics
  - Taking feature similarity into account
  - *Fine tuning* of SVM parameters
  - SVM with AMS_2 metrics
- **SVM threshold optimization**:
  - $AMS_2 \approx 3.45$
  - better than initial ATLAS method *MultiBoost*

Possible Improvements

- Possible improvements:
  - Additional threshold optimization - nearer to AMS metrics
  - Taking feature similarity into account
  - *Fine tuning* of SVM parameters
  - SVM with AMS_2 metrics
- **SVM threshold optimization**:
  - $AMS_2 \approx 3.45$
  - better than initial ATLAS method *MultiBoost*







Conclusions

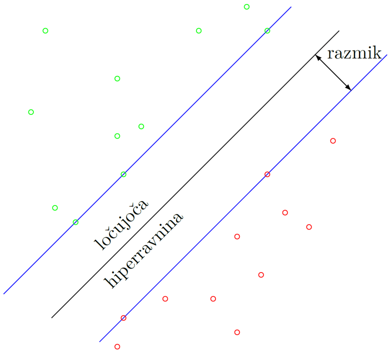
- Introspection into Higgs boson search
- ATLAS detector data analysis
- Machine learning - data engineer point of view
- Better results with SVM than others
- Standard metrics: our method works better than XGBoost; need for optimization
- XGBoost ($AMS_2 \approx 3.8$), MultiBoost ($AMS_2 \approx 3.34$), **opt. SVM (**$AMS_2 \approx 3.45$**)**, GBC ($AMS_2 \approx 2.8$)

Questions?

Thank you for your attention!

SVM

- $H(w, b) = w^T x + b = 0$, normalization $\min |w^T x + b| = 1$, support vectors x_1, x_2
- Gap maximization
- Classifcation function $g(x) = \text{sign}\left(\sum_{i=1}^{n} \alpha y_i K(x^{(i)}, x) + b\right)$
- Kernels:
  - linear: $K(x^{(i)}, x^{(j)}) = x^{(i)T} x^{(j)}$
  - polynomial: $K(x^{(i)}, x^{(j)}) = (\gamma x^{(i)T} x^{(j)} + c)^d$, $\gamma > 0$
  - gaussian (*radial basis function*, RBF): $K(x^{(i)}, x^{(j)}) = \exp(-\gamma |x^{(i)} - x^{(j)}|^2)$, $\gamma > 0$



Razvejitvena razmerja

Razpadni kanal	Razvejitveno razmerje
$H \to \gamma\gamma$	2.28×10^{-3}
$H \to Z^0 Z^0$	2.64×10^{-2}
$H \to W^+ W^-$	2.15×10^{-1}
$H \to \tau^+ \tau^-$	6.32×10^{-2}
$H \to b\bar{b}$	5.77×10^{-1}
$H \to Z^0 \gamma$	1.54×10^{-3}
$H \to \mu^+ \mu^-$	2.19×10^{-4}

Table: Razvejitvena razmerja za Higgsov bozon z maso $m_H = 125\,\text{GeV}$ po standardnem modelu (vir: PDG).

Razpadni kanal	Razvejitveno razmerje
$\tau \to \pi^- \pi^0 \nu_\tau$	2.551×10^{-1}
$\tau \to e^- \bar{\nu}_e \nu_\tau$	1.785×10^{-1}
$\tau \to \mu^- \bar{\nu}_\mu \nu_\tau$	1.736×10^{-1}
$\tau \to \pi^- \nu_\tau$	1.091×10^{-1}
$\tau \to \pi^- \pi^+ \pi^- \nu_\tau$	9.00×10^{-2}
$\tau \to \pi^- \pi^+ \pi^- \pi^0 \nu_\tau$	2.70×10^{-2}

Table: Razvejitvena razmerja najpogostejših razpadnih načinov delca τ (vir: PDG).

SVM - izpeljava

- Maksimiziramo $2/||w||$, kar pomeni, da minimiziramo $||w||$ ali $||w^2||/2$. Robni pogoj $y^{(i)}(w^T x^{(i)} + b) > 1$ za $i = 1, 2, \ldots, n$.
- $L(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i (y^{(i)}(w^T x^{(i)} + b) - 1)$, $y^{(i)} \in \{-1, 1\}$
- $\frac{\partial}{\partial b} L(w, b, \alpha) = 0$ in $\frac{\partial}{\partial w} L(w, b, \alpha) = 0$
- $\sum_{i=1}^{n} \alpha_i y^{(i)} = 0 \qquad \text{in} \qquad w = \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)}$
- $L(w, b, \alpha) = \frac{1}{2}||w||^2 - \overbrace{w^T \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)}}^{||w||^2} - \overbrace{\sum_{i=1}^{n} \alpha_i y^{(i)} b}^{0} + \sum_{i=1}^{n} \alpha_i = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}||w||^2$
- Dualni problem: $D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} \left\langle x^{(i)}, x^{(j)} \right\rangle$, $\alpha_i \geq 0$ in $\sum_{i=1}^{n} \alpha_i y^{(i)} = 0$.
- Minimiziramo $D(\alpha)$, da določimo α_i.

SVM - mehki razmik in jedra

Mehki razmik:

- Minimiziramo $||w||^2 + C\sum_{i=1}^{n}\xi_i$, pri čemer velja robni pogoj $y^{(i)}(w^T x^{(i)} + b) > 1 - \xi_i$, za $i = 1, 2, ..., n$, kjer je ξ označuje mero za napačno klasifikacijo, C pa je parameter, ki nadzira vpliv ξ.

Jedra:

- Formulacija dualnega problema s pomočjo $W(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j K(x^{(i)}, x^{(j)})$.

Pospeševanje (boosting)

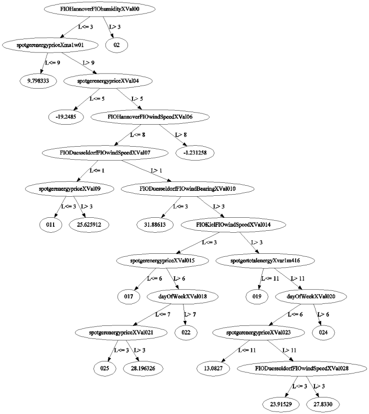
$$
\begin{aligned}
g_0(x) &= \gamma_0 && (1)\\
\gamma_0 &= \arg\min_{\gamma} \sum_{i=1}^{n} L(y^{(i)}, \gamma). && (2)\\
r_m^{(i)} &= -\left[\frac{\partial L(y^{(i)}, g(x^{(i)}))}{\partial g(x^{(i)})}\right]_{g(x)=g_{m-1}(x)} && (3)\\
\mathcal{D}_m &= \{(x^{(i)}, r_m^{(i)})\}_{i=1}^{n} \rightarrow h_m(x) && (4)\\
g_m(x) &= g_{m-1}(x) + \gamma_m h_m(x) && (5)\\
\gamma_m &= \arg\min_{\gamma} \sum_{i=1}^{n} L(y^{(i)}, g_{m-1}(x^{(i)}) + \gamma h_m(x^{(i)})) && (6)
\end{aligned}
$$

Logistična regresija

- Pri linearni regresiji minimiziramo $J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$
- Pri logistični regresiji namesto sumanda uvedemo stroškovno funkcijo $\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{če je } y = 1 \\ -\log(1 - h_\theta(x)) & \text{če je } y = 0 \end{cases}$
- Ker je $y \in \{0, 1\}$, sledi ...
- $J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))\right]$
- Gradientni spust $\rightarrow \frac{\partial}{\partial x_i} J(\theta)$
- Korak iteracije $\theta_j := \theta_j - \alpha\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$

Odločitvena drevesa

- Entropija: $S(\mathcal{D}) = -p_+ \log_2 p_+ - p_- \log_2 p_-$
- Informacijski pribitek: $G(\mathcal{D}, Z) = S(\mathcal{D}) - \sum_{v \in \mathcal{Z}_Z} \frac{|\mathcal{D}_v|}{|\mathcal{D}|} S(\mathcal{D}_v)$, kjer je Z izbrana značilka, $\mathcal{Z}_Z$ množica vrednosti značilke Z, $\mathcal{D}_v$ pa podmnožica $\mathcal{D}$ s primerki z vrednostjo parametra $Z = v$.



Lagrange / Hamilton

- $\frac{\partial L}{\partial x_k} - \frac{d}{dt}\frac{\partial L}{\partial \dot{x}_k} = 0$
- Delec v EM polju: $L = \frac{1}{2}m\dot{\mathbf{r}}^2 - e\varphi + e\mathbf{vA}$
- $H = T + V; H = H(q, p, t)$
- $\frac{dp}{dt} = -\frac{\partial H}{\partial q}$
- $\frac{dq}{dt} = +\frac{\partial H}{\partial p}$

Maxwell in ostale uporabne zveze

$$
\begin{aligned}
\nabla \cdot E &= \frac{\rho}{\epsilon_0} \\
\nabla \cdot B &= 0 \\
\nabla \times E &= -\frac{\partial B}{\partial t} \\
\nabla \times B &= \mu_0 j + \epsilon_0 \mu_0 \cdot \frac{\partial E}{\partial t} \\
B &= \nabla \times A \\
E &= -\nabla \varphi - \frac{\partial A}{\partial t} \\
\frac{\partial \rho}{\partial t} + \nabla j &= 0
\end{aligned}
$$