Basics of Information Theory

for machine learning

$P(A)$

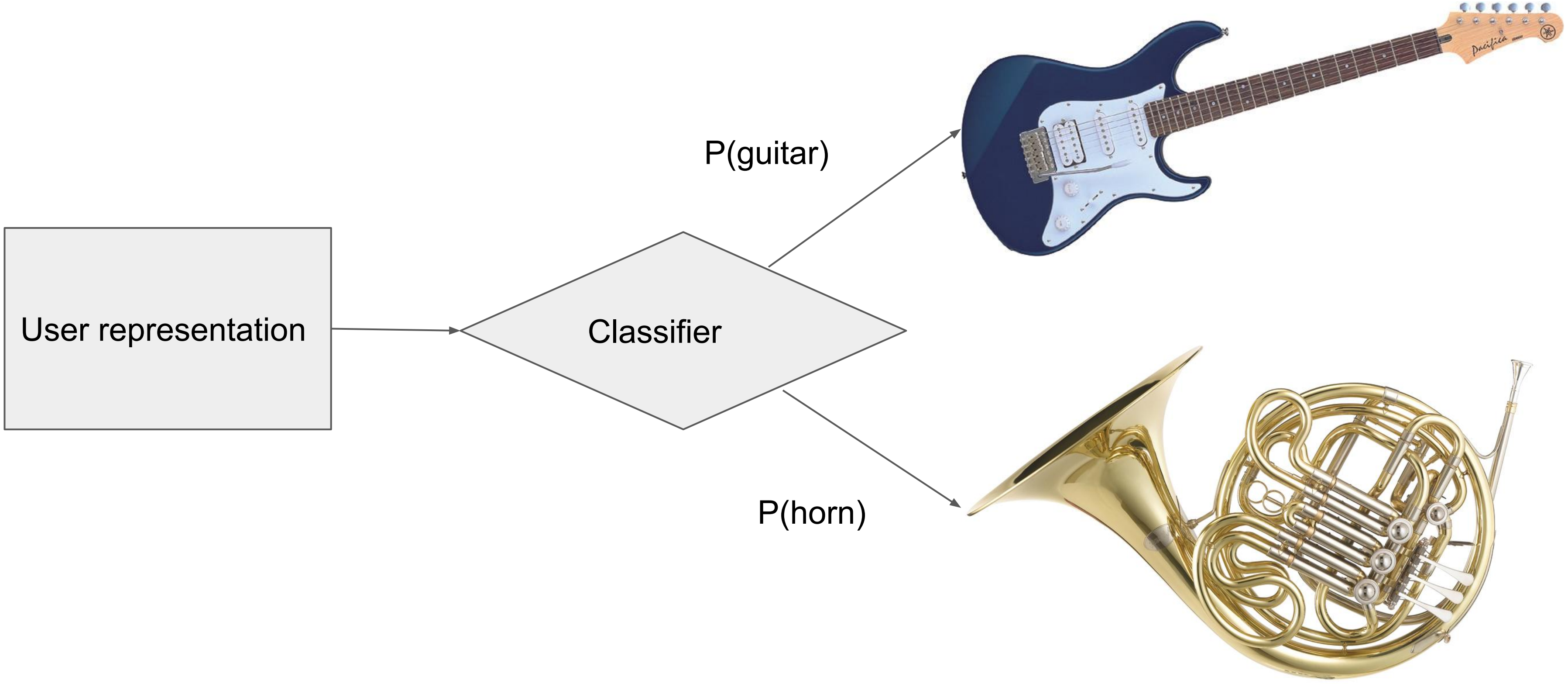
Frequentist interpretation

Image from: http://www.shmoop.com/basic-statistics-probability/probability.html

Frequentist interpretation

$$P(x) = \lim_{n_t \to \infty} \frac{n_x}{n_t}.$$

Bayesian interpretation



Bayesian interpretation

a reasonable expectation

that represents the state of knowledge

Sources of uncertainty

Image from: http://braveleaps.com/wp-content/uploads/2013/10/Uncertainty2.jpg

Inherent stochasticity

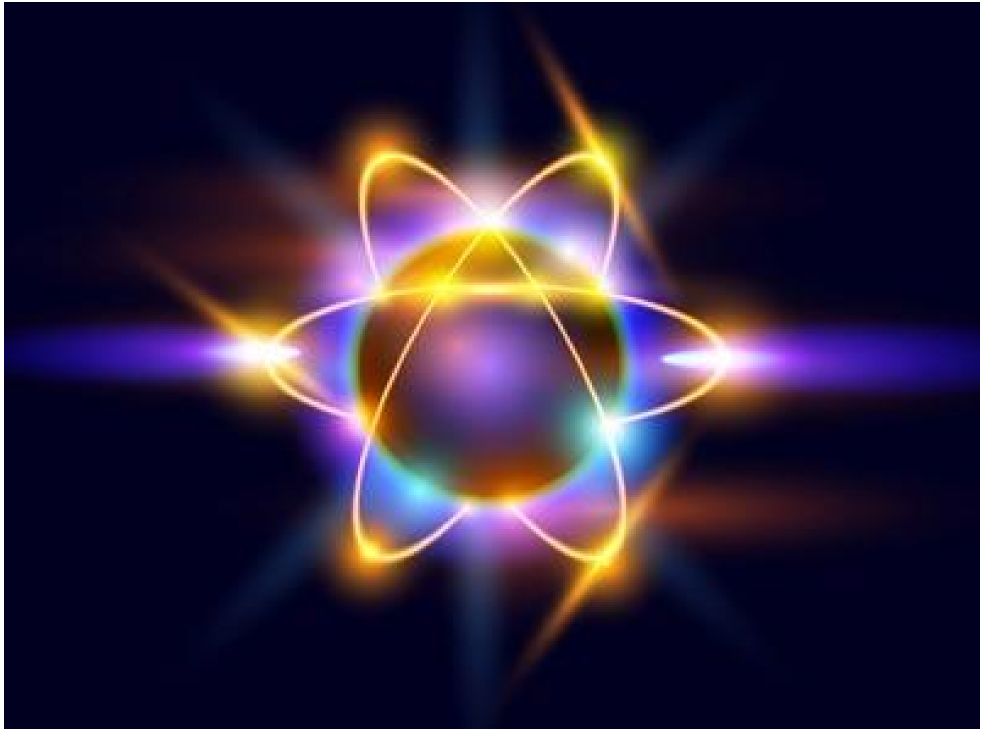


Image from: http://www.science4all.org/wp-content/uploads/2013/01/atom711.jpg

Incomplete observability

Image from:
https://img.clipartfest.com/b4b32c0549e64021f6408edecb1296be_from-clipartcom-behind-the-tree-clipart

Model limitations

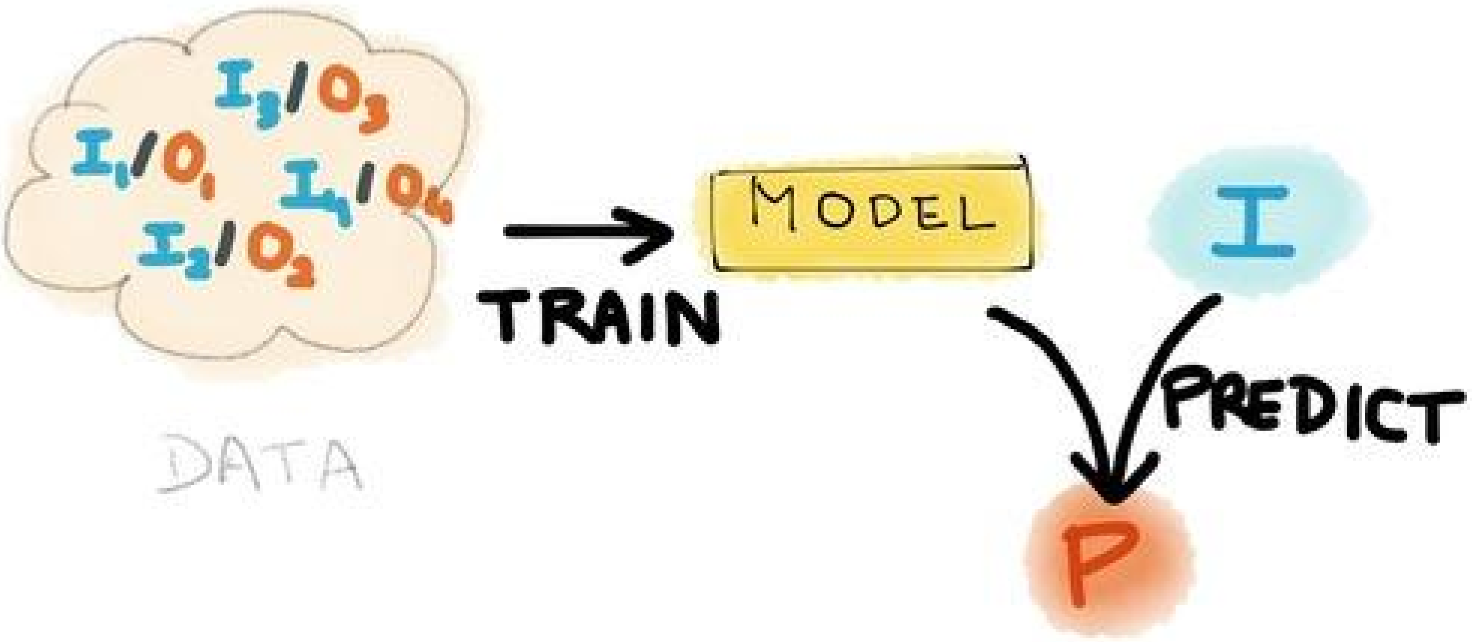


Image from:
http://static1.squarespace.com/static/5206b718e4b0bdc26006bae2/t/554771abe4b0deabe55d5faa/143074

Probability Mass Function

- The domain of P must be the set of all possible states of x.
- $\forall x \in x, 0 \leq P(x) \leq 1$.
- $\Sigma x \in x P(x) = 1$.

Uniform distribution

$$P\ (x = x_i) = 1/k$$

Zipfian distribution

$$f(k;s,N)=\frac{1/k^s}{\sum_{n=1}^{N}(1/n^s)}$$

Zipf's Law

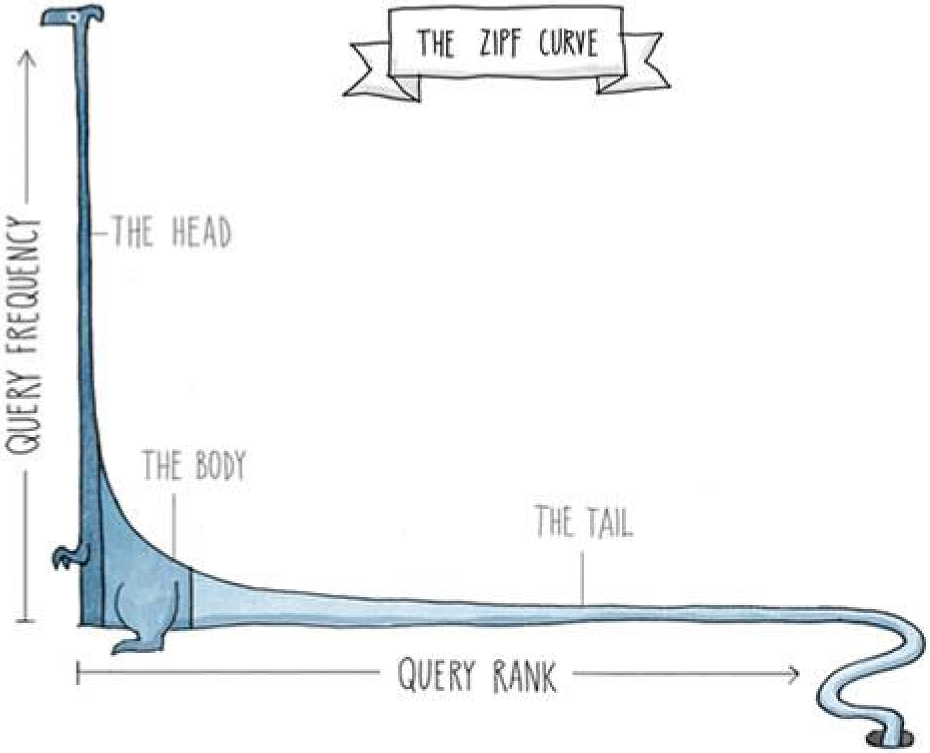


Image from: https://i.stack.imgur.com/qSUgV.jpg

Additive smoothing

$$\hat{\theta}_i = \frac{x_i + \alpha}{N + \alpha d} \qquad (i = 1, \ldots, d)$$

Open world

P(president | ”Obama”) = 0.95

P(town | ”Obama”) = 0.02

P(something_else | ”Obama”) = 0.03

More meanings at: https://en.wikipedia.org/wiki/Obama_(disambiguation)

Joint probability

	King	Ace
Diamonds	1/4	1/4
Spades	1/4	1/4

Marginal probability

	King	Ace	
Diamonds	1/4	1/4	1/2
Spades	1/4	1/4	1/2
	1/2	1/2	

Conditional probability

$$p(y|x)=p(x,y)/p(x)$$

$$p(x,y)=p(x)\cdot p(y|x)$$

Independence ($x \perp y$)

$$p(x, y) = p(x)p(y)$$

$$p(x|y) = p(x)$$

Independence (cont’d)

Images from: shop.fender.com

Dependent variables

	Red	Blonde	
Stratocaster	1/2	1/8	5/8
Telecaster	1/8	1/4	3/8
	5/8	3/8	

Dependent variables

$$p(y|x)=p(x,y)/p(x)$$

$P(\text{red}) = \frac{5}{8} = 0.625$

$P(\text{stratocaster}) = \frac{5}{8}$

$P(\text{red},\text{stratocaster}) = \frac{1}{2}$

$P(\text{red}|\text{stratocaster}) = \frac{1}{2} / \frac{5}{8} = 0.8$

Dependent variables

$$p(y|x)=p(x,y)/p(x)$$

$P(\text{blonde}) = \frac{3}{8} = 0.375$

$P(\text{telecaster}) = \frac{3}{8}$

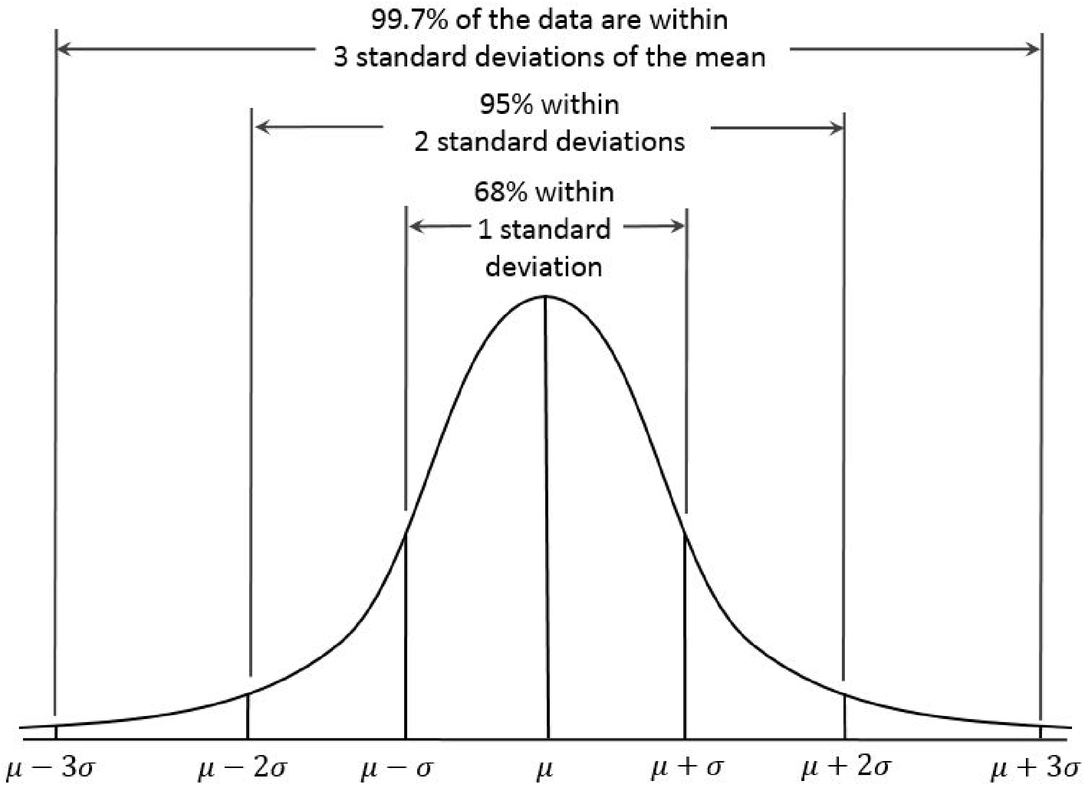
$P(\text{blonde},\text{telecaster}) = \frac{1}{4}$

$P(\text{blonde}|\text{telecaster}) = \frac{1}{4} / \frac{3}{8} = \frac{2}{3} \approx 0.(6)$

Bayes rule

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Gaussian distribution



Gaussian distribution

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Expectation

$$\mathbb{E}_{\mathrm{x}\sim P}[f(x)] = \sum_{x} P(x)f(x),$$

Expectations are linear

$$\mathbb{E}_{\mathbf{x}}[\alpha f(x) + \beta g(x)] = \alpha \mathbb{E}_{\mathbf{x}}[f(x)] + \beta \mathbb{E}_{\mathbf{x}}[g(x)]$$

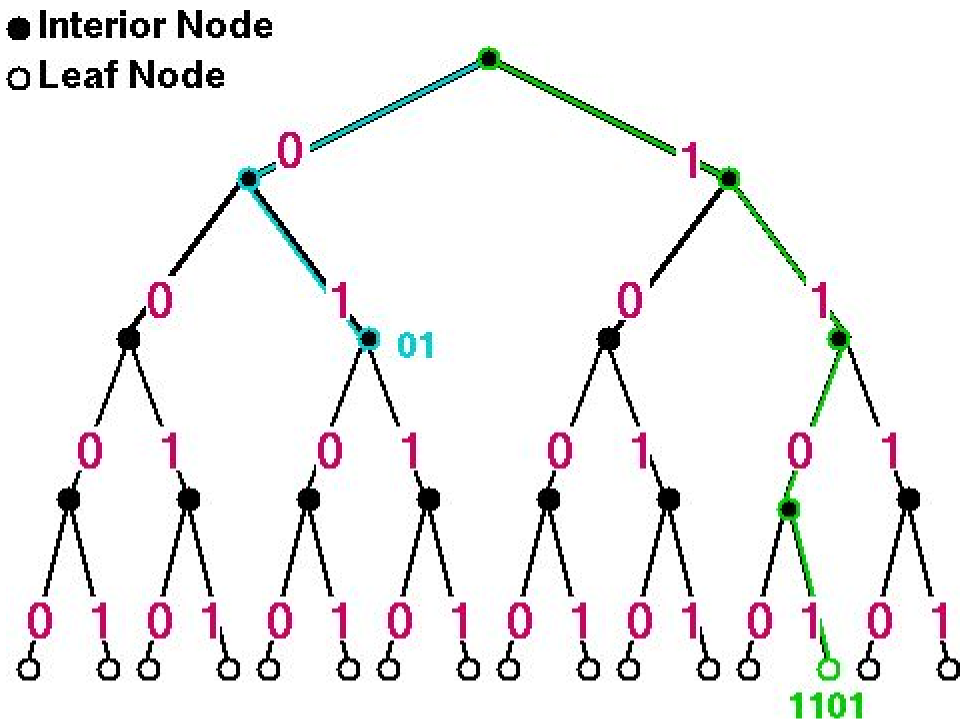
Variance

$$\mathrm{Var}(f(x)) = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right]$$

Covariance

$$\mathrm{Cov}(f(x), g(y)) = \mathbb{E}\left[\left(f(x) - \mathbb{E}\left[f(x)\right]\right)\left(g(y) - \mathbb{E}\left[g(y)\right]\right)\right]$$

Binary encoding



Binary encoding

x bits may represent 2^x equally likely events

z equally likely events may be represented by $\log_2 z$ bits

Other base(s)

x nats may represent e^x equally likely events

z equally likely events may be represented by ln(z) nats

Huffman coding

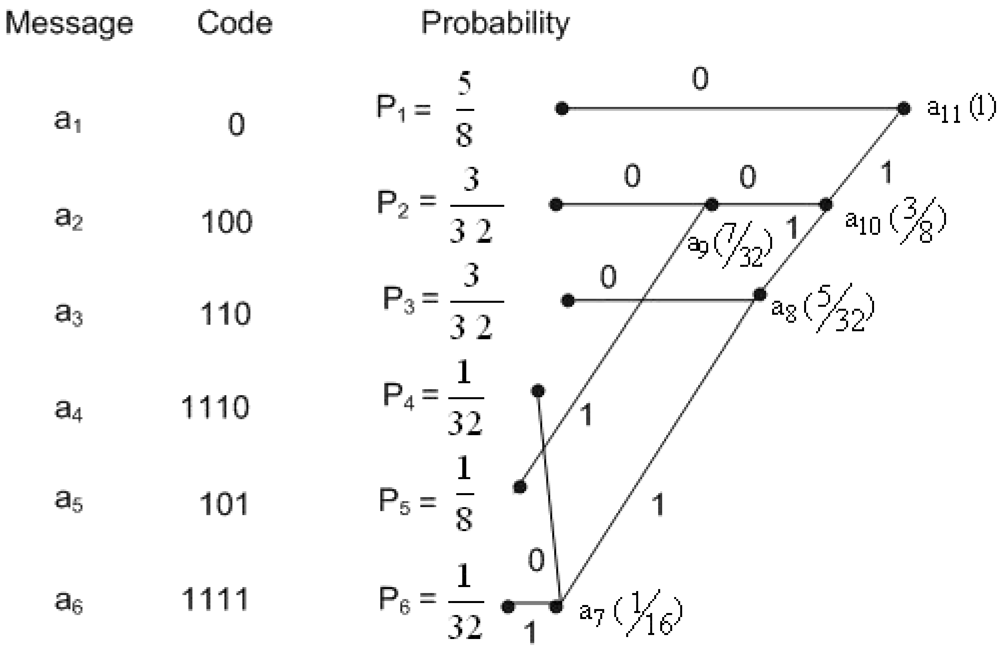
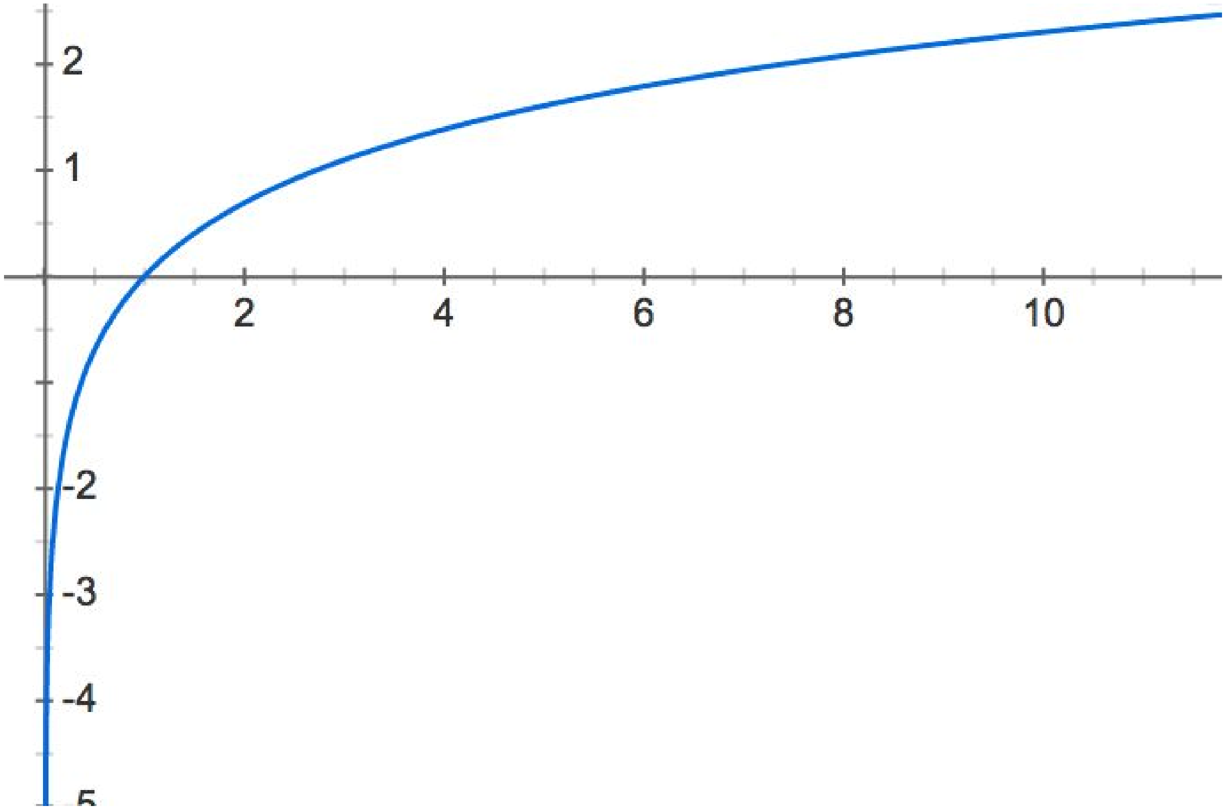


Image from: http://nptel.ac.in/courses/117104069/chapter_7/fig12b.gif

Self-information

$$I(x) = -\log P(x)$$

ln(x)



Shannon Entropy

$$H(\mathrm{x}) = \mathbb{E}_{\mathrm{x} \sim P}[I(x)] = -\mathbb{E}_{\mathrm{x} \sim P}[\log P(x)]$$

Cross-entropy

$$H(P, Q) = -\mathbb{E}_{\mathrm{x} \sim P} \log Q(x)$$

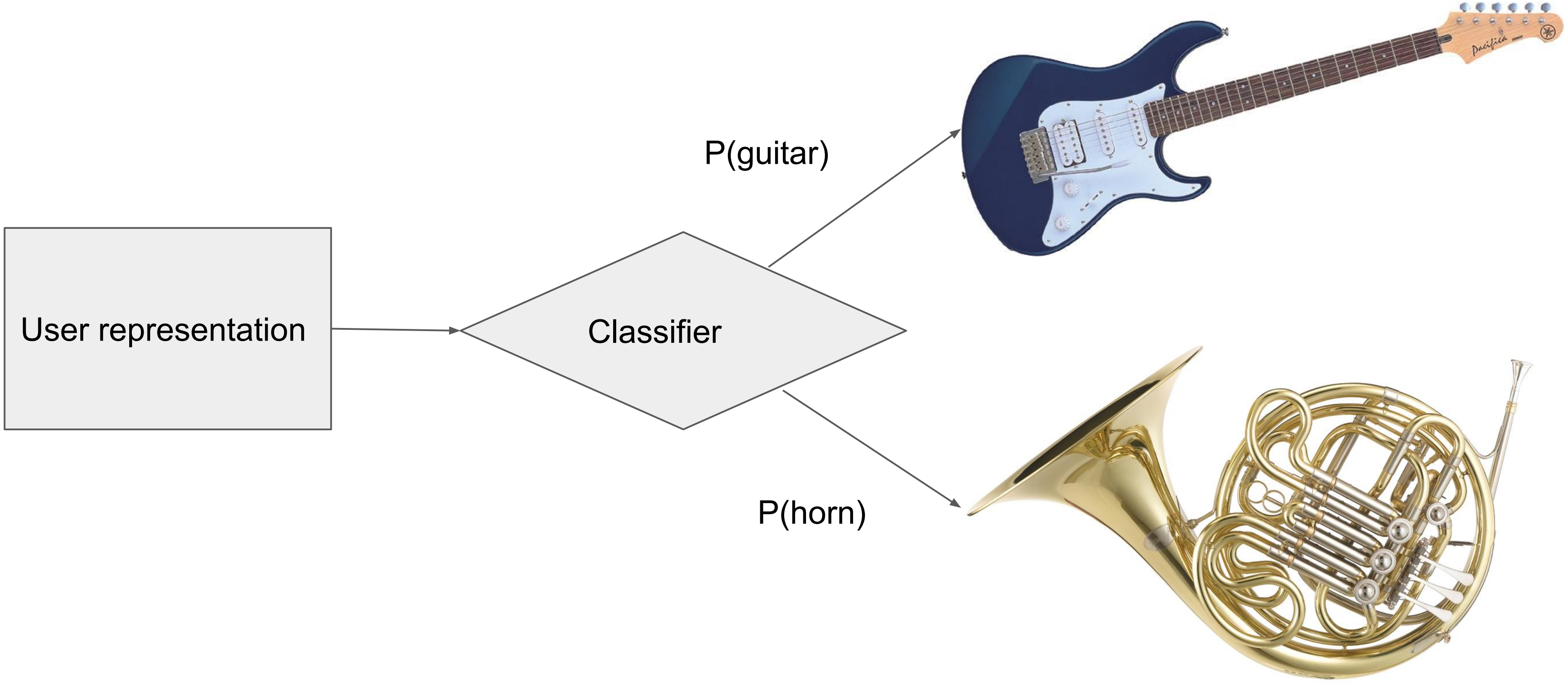
Cross-entropy

$$H(p,q) = -\sum_{x} p(x)\, \log q(x)$$

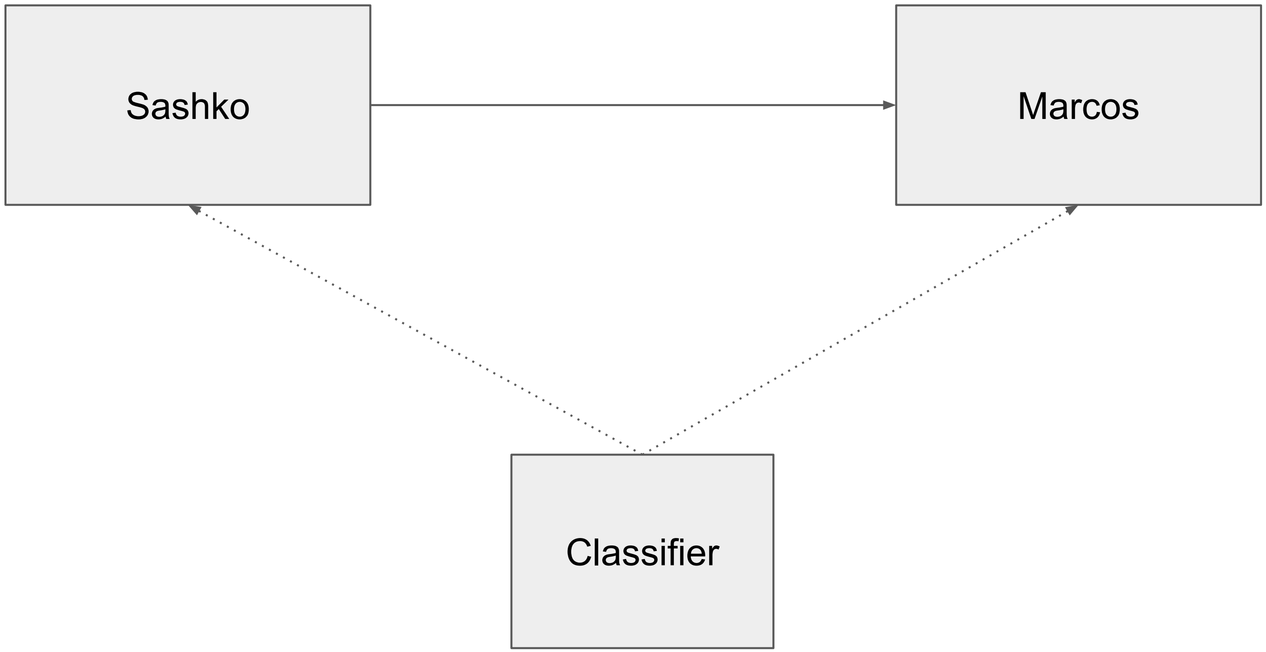
Logistic loss

$$- y \log \hat{y} - (1 - y) \log(1 - \hat{y})$$

Back to classification



Classification



Mutual information

$$I(X,Y)=H(X)+H(Y)-H(X,Y)$$

PMI

$$\operatorname{pmi}(x;y) \equiv \log \frac{p(x,y)}{p(x)p(y)}$$

Distributed representation

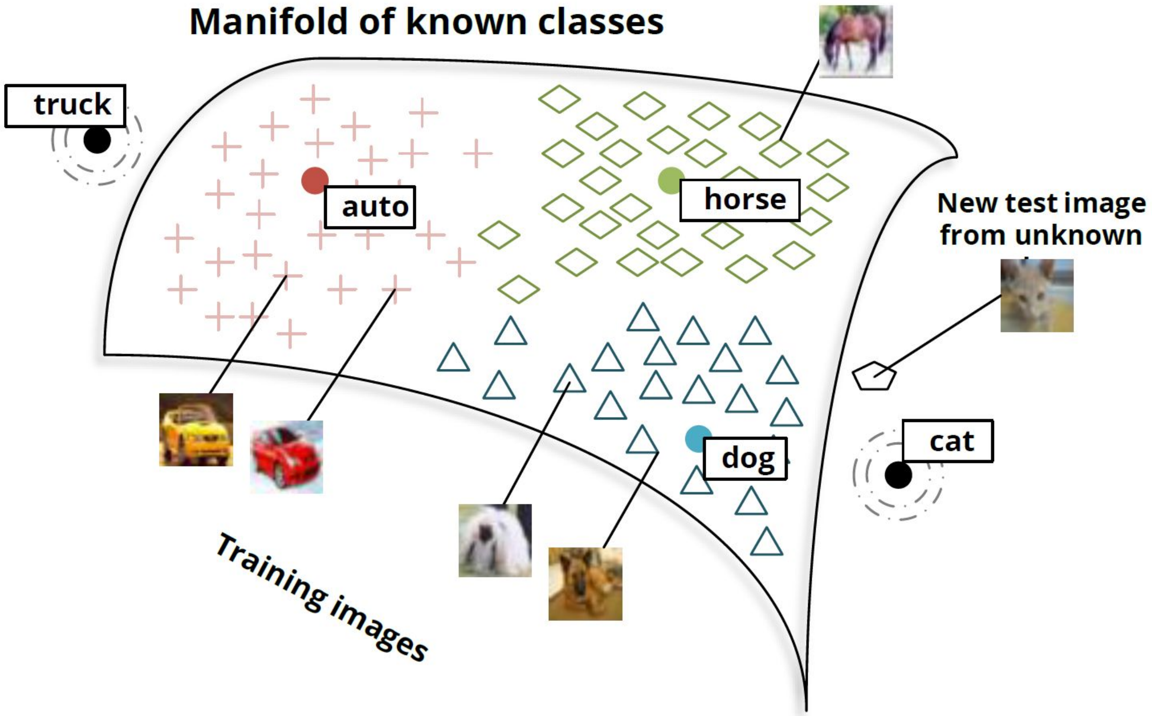


Image from:
http://colah.github.io/posts/2014-07-NLP-RNNs-Representations/img/Socher-ImageClassManifold.png

Further reading: probability and statistics

1. Probability Theory: The Logic of Science (by E. T. Jaynes)
   http://bayes.wustl.edu/etj/prob/book.pdf
2. Introduction to Statistical Thinking (by Benjamin Yakir)
   http://pluto.huji.ac.il/~msby/StatThink/IntroStat.pdf
3. Introduction to Probability and Statistics Using R (by G. Jay Kerns)
   https://cran.r-project.org/web/packages/IPSUR/vignettes/IPSUR.pdf

Further reading: Information Theory

1. Information Theory, Inference, and Learning Algorithms (by David MacKay)
   http://www.inference.phy.cam.ac.uk/itprnn/book.html
2. Elements of information theory (by Joy A. Thomas and Thomas M. Cover)
   http://www.di-srv.unisa.it/professori/uv/TI2/libro.pdf
3. Visual Information Theory (by Christopher Olah)
   http://colah.github.io/posts/2015-09-Visual-Information/
4. Information Theory: A Tutorial Introduction (by James V. Stone)
   http://jim-stone.staff.shef.ac.uk/BookInfoTheory/

Further reading: Machine Learning

1. Deep Learning (by Ian Goodfellow, Yoshua Bengio, and Aaron Courville)
   https://www.deeplearningbook.org/

Thanks!