SJSU SAN JOSÉ STATE UNIVERSITY

DATA 220: Mathematical Methods for Data Analytics

Dr. Mohammad Masum

Optimization (Cont.)

- Batch Gradient Descent
- Stochastic Gradient Descent
- Mini-Batch Gradient Descent

Optimization (Cont.)

- Batch Gradient Descent
  - All samples in one iteration ➔ iteration = epoch

$$\theta = \theta - \eta \,.\Delta_\theta J\,(\theta)$$

Optimization (Cont.)

- Stochastic Gradient Descent
  - One random sample at each iteration
  - Performs weights update after each iteration for each sample
  - Number of iteration per epoch = number of samples

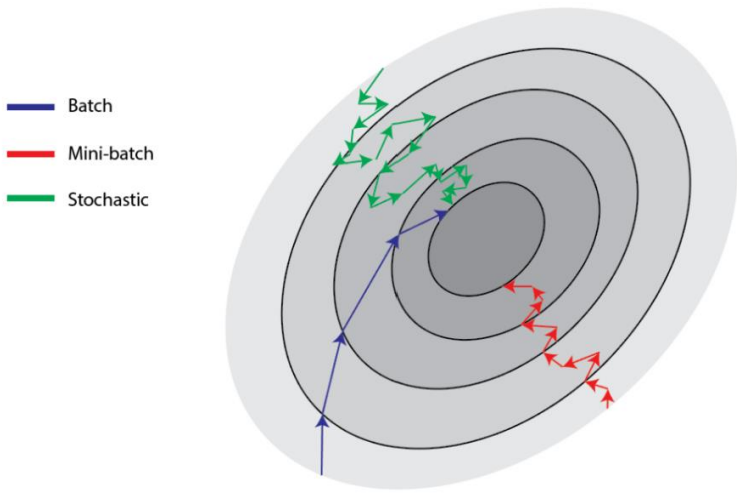
$$\theta = \theta - \eta \,.\, \Delta_\theta J\,(\theta; X^i, y^i)$$

https://www.baeldung.com/cs/learning-rate-batch-size

Optimization (Cont.)

- Mini Batch Gradient Descent
  - Take advantage of both batch and stochastic gradient descent
  - weights update after passing one batch in one iteration
  - Number of iteration per epoch = number of samples/batch size

$$\theta = \theta - \eta \,.\, \Delta_\theta J\,(\theta; X^{i:i+n}, y^{i:i+n})$$



https://www.baeldung.com/cs/learning-rate-batch-size

Information Theory

- Information Theory
  - Quantification, storage, and communication of information – provides a mathematical framework for understanding the fundamental limits of data compression and transmission
  - Information can be defined as the reduction of uncertainty
    - In DS: deal with large amounts of data, and the goal is to extract meaningful information from it

Shannon's Information Theory - is a mathematical framework for studying the transmission of information over a communication channel. The theory was introduced by Claude Shannon in 1948, and it has become a fundamental concept in information science, computer science, and electrical engineering.

https://medium.com/udacity/shannon-entropy-information-gain-and-picking-balls-from-buckets-5810d35d54b4

Information Theory

Reprinted with corrections from *The Bell System Technical Journal,*
Vol. 27, pp. 379–423, 623–656, July, October, 1948.

A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

Information Theory

- Shannon's information theory- Entropy quantifies the amount of uncertainty (surprise) involved in a message (value of a random variable)
  - Use cases in ML: Decision Tree, Feature Selection, Comparing probability distribution (VAE), loss function in classification task, Active ML..

The more random and unpredictable the data is, the higher the entropy.

> Why unify information theory and machine learning? Because they are two sides of the same coin. [...] Information theory and machine learning still belong together. Brains are the ultimate compression and communication systems. And the state-of-the-art algorithms for both data compression and error-correcting codes use the same tools as machine learning.
>
> — Page v, *Information Theory, Inference, and Learning Algorithms*, 2003.

Information Theory – Calculate information for an event

- Recall - Expected Value (mean) of a random variable (Discrete)
  - $E(X) = \sum_{all\ possible\ x} x\ P(X = x)$
  - Example: the expected value of a roll of a die:

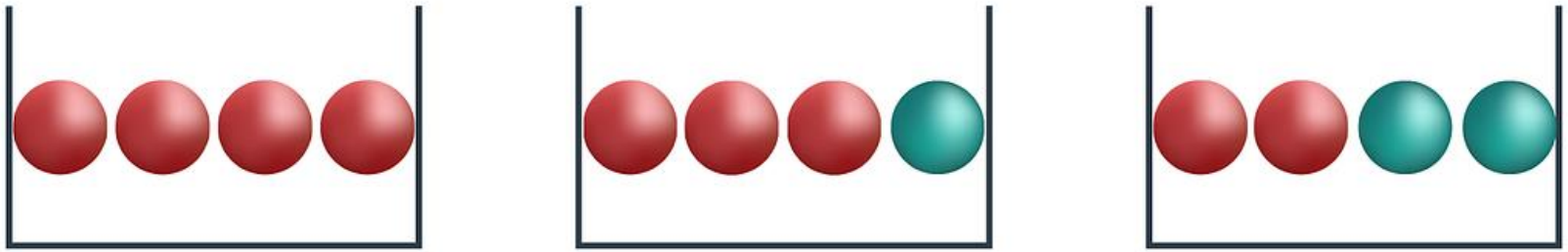
$$1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) = 3.5$$

Information Theory – Calculate information for an event

- Quantifying information - measuring how much surprise is in an event
- Probability of an event can be used for calculate the amount of information (surprise) of an event
  - inversely related to probability of an event

    Low probability event – more surprising - high information

    High probability event – unsurprising - low information



High Knowledge
Low Entropy

Medium Knowledge
Medium Entropy

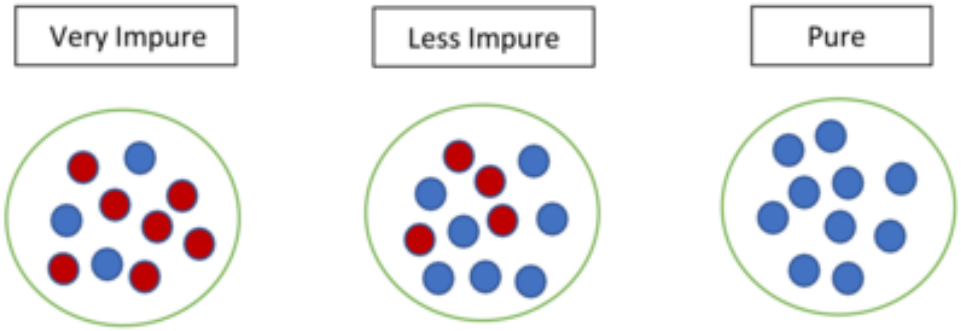
Low Knowledge
High Entropy

Slide credit - StatQuest/Entropy
https://medium.com/udacity/shannon-entropy-information-gain-and-picking-balls-from-buckets-5810d35d54b4

Information Theory – Entropy

For Discrete Random Variable X with PMF P(X), Entropy:

$$H(X) = -\sum_{x \in X} P(x) \log(P(x))$$



The leftmost figure is very impure and has high entropy corresponding to higher disorder and lower information value. As we go to the right, the entropy decreases, and the information value increases.

https://en.wikipedia.org/wiki/Information_gain_(decision_tree)

Information Theory – Entropy

- Specific conditional entropy $\boldsymbol{H(y|\, X = v)}$
  - Entropy of y among only for those records where $X = v$
- Conditional entropy, $H(y|X)$
  - expected value of specific conditional entropy
  - $\boldsymbol{H(y \mid X) = \sum p(X = v)\, H(y \mid X = v)}$

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

https://www.youtube.com/watch?v=wNhnAogeHJQ

Information Theory – Information Gain

- Information gain
  - The amount of information gain about a random variable from observing another random variable
  - IG can tell us how much information about y (target variable) is contained in X
  - $\mathbf{IG}\ (\mathbf{y} \mid \mathbf{X}) = \boldsymbol{H}(\boldsymbol{y}) - \boldsymbol{H}\ (\boldsymbol{y} \mid \boldsymbol{X})$
  - We can find the IG for each of the variable of a dataset and rank
  - We want features with high information gain that tells us lot about the value of y
- Find most important features from this dataset:

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

https://nulpointerexception.com/2017/12/16/a-tutorial-to-understand-decision-tree-id3-learning-algorithm/

Information Theory – Gini Index

- Measure of Impurity:
  - Gini Index
    - Most commonly used to measure inequality of income or wealth
    - Gini index for a given node *t:*

$$GINI(t) = 1 - \sum_j [p(j|t)]^2$$

$p(j|t)$ is a conditional probability, can be measured by **relative frequency of class j at node t**

- ***Gini Index = 0*** when all data belong to single class only
- ***Gini index = 0.5*** when all data are equally distributed

C1	0
C2	6
Gini=0.000	

C1	3
C2	3
Gini=0.500	

Information Theory – Gini Index

- Calculate impurity by *Gini Index*

$$GINI(t) = 1 - \sum_j [p(j|t)]^2$$

C0: 9 C1: 1

C0: 5 C1: 5

Information Theory – Gini Index

- Calculate impurity of a node by ***Gini Index***

$$GINI(t) = 1 - \sum_j [p(j|t)]^2$$

C0: 9
C1: 1

Gini index = 1 – (9/10)^2 – (1/10)^2
= 0.18

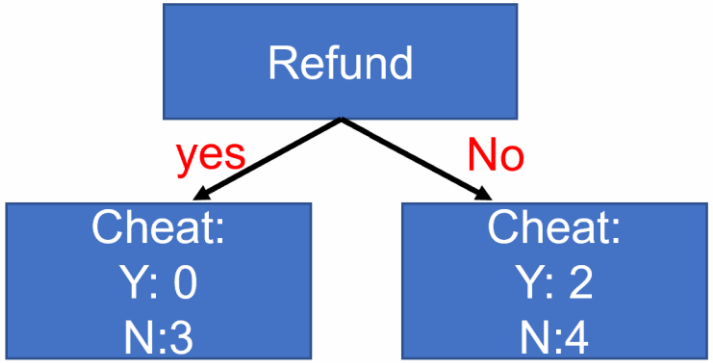
C0: 5
C1: 5

Gini index = *1 – (5/10)^2 – (5/10)^2*
= ***0.5***

$$GINI(t) = 1 - \sum_j [p(j|t)]^2$$

Information Theory – Gini Index

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



$$GI = 1 - \left(\frac{0}{3}\right)^2 - \left(\frac{3}{3}\right)^2 = \mathbf{0}$$

$$GI = 1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2 = \mathbf{0.49}$$

$$\text{Gini } Index\ for\ Refund = \frac{3}{10} * 0 + \frac{7}{10} * 0.49 = \mathbf{0.34}$$

Information Theory – Divergence between Probability Distribution

- Quantify the difference between probability distributions for a random variable
  - Kulback-Leibler Divergence

Information Theory – KL Divergence

- Kullback-Leibler Divergence
  - Measure how one probability distribution *p* differs from the second *q* (is not distance between two distribution)
  - Is not symmetric (meaning that divergence from *p* to q distribution and divergence from *q* to *p* distribution is not identical) $\boldsymbol{D_{KL}(p||q) \neq D_{KL}(q \,||\, p)}$
  - Interpretation: $\boldsymbol{D_{KL}(p||q)}$ – is the information grain achieved when distribution p would be used instead of distribution *q*

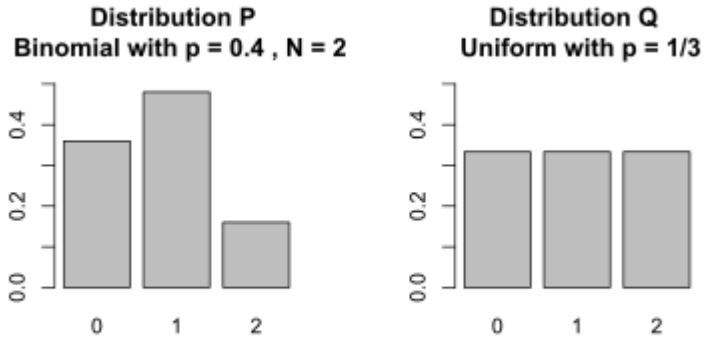
For discrete random variables

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right).$$

For continuous random variables

$$D_{\mathrm{KL}}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$$

Information Theory – KL Divergence

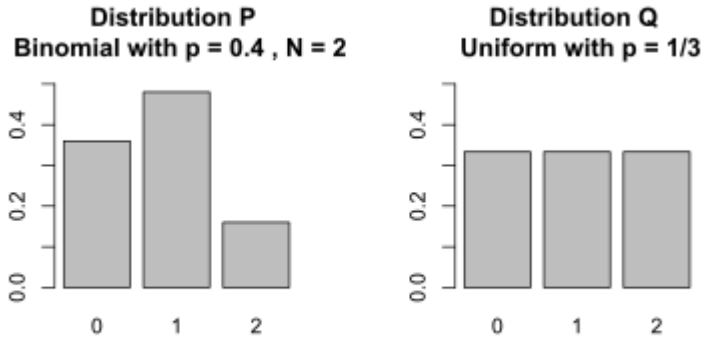


x	0	1	2
Distribution $P(x)$	$\frac{9}{25}$	$\frac{12}{25}$	$\frac{4}{25}$
Distribution $Q(x)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence

Information Theory – KL Divergence

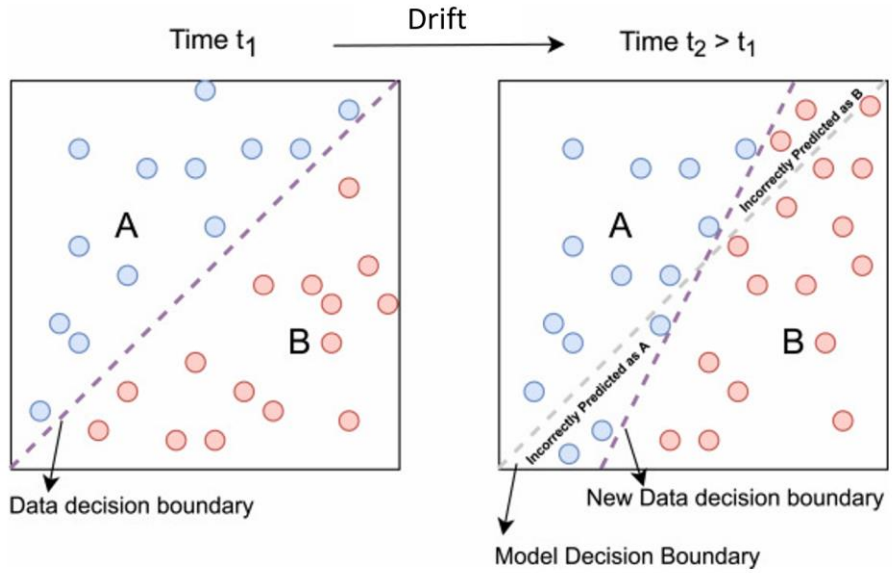
$$
\begin{aligned}
D_{\mathrm{KL}}(P \parallel Q) &= \sum_{x \in \mathcal{X}} P(x) \ln\left(\frac{P(x)}{Q(x)}\right) \\
&= \frac{9}{25} \ln\left(\frac{9/25}{1/3}\right) + \frac{12}{25} \ln\left(\frac{12/25}{1/3}\right) + \frac{4}{25} \ln\left(\frac{4/25}{1/3}\right) \\
&= \frac{1}{25}\left(32\ln(2) + 55\ln(3) - 50\ln(5)\right) \approx 0.0852996 \\
D_{\mathrm{KL}}(Q \parallel P) &= \sum_{x \in \mathcal{X}} Q(x) \ln\left(\frac{Q(x)}{P(x)}\right) \\
&= \frac{1}{3} \ln\left(\frac{1/3}{9/25}\right) + \frac{1}{3} \ln\left(\frac{1/3}{12/25}\right) + \frac{1}{3} \ln\left(\frac{1/3}{4/25}\right) \\
&= \frac{1}{3}\left(-4\ln(2) - 6\ln(3) + 6\ln(5)\right) \approx 0.097455
\end{aligned}
$$



x	0	1	2
Distribution $P(x)$	$\frac{9}{25}$	$\frac{12}{25}$	$\frac{4}{25}$
Distribution $Q(x)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Data Drift – Challenge of Changing Data Over Time

- Data Drift –
  - Dynamic evolution of statistical properties in datasets
  - Essential for ensuring ongoing performance and reliability of predictive models
  - Reduce time and resources for model maintenance
- Scenarios of Data Drift –
  - E-commerce: User behavior changes
  - Healthcare: Changes in Patient data
- Data Drift Detection –
  - Statistical tests: Compare data distribution – KL Divergence
  - Model based approaches: assess models' performance



Information Theory – Cross Entropy

- Cross Entropy – Quantifies the difference between two probability distributions
- Mathematically: $H(P, Q) = -\sum_{x \in X} P(x) \log\ (Q(x))$