

Evaluation of Gold Labels

Evaluation of Gold Labels

Height (inches)	Weight (kgs)	B.P.Sys	B.P.Dia	Heart disease	
\vec{x}				y	$\hat{y} = h(\vec{x})$
62	70	120	80	No	No
72	90	110	70	No	Yes
74	80	130	70	No	No
65	120	150	90	Yes	Yes
67	100	140	85	Yes	No
64	110	130	90	No	Yes
69	150	170	100	Yes	Yes

- y : Gold standards / Gold Labels / Ground Truth
- \hat{y} : Predicted Labels

A lot is at stake when we deploy a classifier...

Can we trust Gold Labels?

(How) can we trust the Gold Labels?

❑ **Objective Data:** Gathered from the real world


- Measurements of heights, weights, weather phenomena
 - Aberrations, sensor malfunctions and errors

❑ **Subjective Data:** Annotated by humans (experts or community)

- Tagging emotions, Spam / Ham, Sentiment, misinformation ...
 - Errors of judgement, biases, human error

❑ **Use multiple sources for each label**

- Multiple sensors measuring the same phenomenon
- Multiple humans annotating the same data
 - Inter-annotator agreements
- Resource intense
 - Partial overlap of data



Labeling is not
very easy...

❑ **Is annotator agreement the answer to all our worries?**

- Distribution of annotators
- Chance agreement

Gold Labels: Annotators and Agreement

Multiple Annotators

Instances 0	A_1	A_2	Raw (Observed) Agreement	A_{Rand}	Agreement (A_1, A_{Rand})	Agreement (A_2, A_{Rand})
1	Positive	Positive	1	Negative	0	0
2	Negative	Negative	1	Positive	0	0
3	Positive	Negative	0	Positive	1	0
4	Negative	Positive	0	Negative	1	0
...
			\sum Raw Agreements		\sum Chance Agreements	\sum Chance Agreements

Inter-rater / Inter-annotator agreement

$$\text{Cohen's Kappa } (\kappa) = \frac{P_o - P_e}{1 - P_e} = 1 - \frac{1 - P_o}{1 - P_e}$$

Binary-class
Labels

$$\text{Krippendorff's alpha } (\alpha) = 1 - \frac{D_o}{D_e}$$

Multi-class
Labels

Cohen's Kappa

Suppose that two people are annotating 50 instances as “Yes” or “No”. Suppose disagreement count data was as follows.

		A_2	
		Yes	No
A_1	Yes	$a = 20$	$b = 5$
	No	$c = 10$	$d = 15$

- The observed proportionate agreement is:

$$p_o = \frac{a + d}{a + b + c + d} = \frac{20 + 15}{50} = 0.7$$

- To calculate p_e (the probability of random agreement), we note that:
 - Annotator A_1 said “Yes” for 25 labels. Thus, A_1 said “Yes” 50% of the time.
 - Annotator A_2 said “Yes” for 30 labels. Thus, A_2 said “Yes” 60% of the time.

$$p_{Yes} = \frac{a + b}{a + b + c + d} \cdot \frac{a + c}{a + b + c + d} = 0.5 \times 0.6 = 0.3$$

$$p_{No} = \frac{c + d}{a + b + c + d} \cdot \frac{b + d}{a + b + c + d} = 0.5 \times 0.4 = 0.2$$

$$p_e = p_{Yes} + p_{No} = 0.3 + 0.2 = 0.5$$

$$\text{Cohen's Kappa } (\kappa) = \frac{0.7 - 0.5}{1 - 0.5} = 0.4$$

Useful Resources

- ❑ [Cohen's Kappa. Understanding Cohen's Kappa coefficient | by Kurtis Pykes | Towards Data Science](#)
- ❑ [Computing Krippendorff's Alpha-Reliability \(upenn.edu\)](#)

Calculating Feature Importance

FEATURE SELECTION METHODS

Feature Selection

Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rainy	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

Which Features are important?

Step 1: Find Entropy of whole dataset.

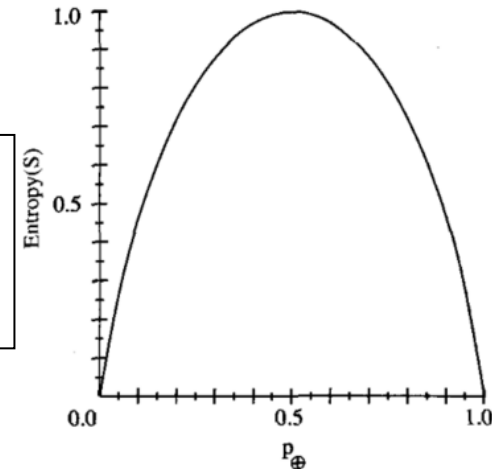
$$Entropy(S) \equiv \sum_{i=1}^c -p_i \log_2 p_i$$

$$Entropy(S) \equiv -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$

$$Entropy(S) = -\frac{9}{14} \times \log_2 \frac{9}{14} - \frac{5}{14} \times \log_2 \frac{5}{14}$$

When Entropy is 1 (maximum) and when 0 (minimum)?

- Entropy is 0 if all members of S belong to the same class.
- Entropy is 1 when S contains equal number of positive and negative examples.



- The maximum value of entropy is $\log(k)$ where k is the number of categories.
- In case of two classes, it's $\log(2)$ which is 1, and in case of 3 classes, it would be $\log(3)$.

Information Gain

Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rainy	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

Information Gain measures the expected *Reduction in Entropy*

$$S = [9+, 5-]$$

$$Entropy(S) = -\frac{9}{14} \times \log_2 \frac{9}{14} - \frac{5}{14} \times \log_2 \frac{5}{14} = 0.94$$

Step 2: Find Entropy for each possible value of each attribute

Attribute: Outlook

Values (Outlook) = Sunny, Overcast, Rain

$$S_{sunny} = [2+, 3-]$$

$$Entropy(S_{sunny}) = -\frac{2}{5} \times \log_2 \frac{2}{5} - \frac{3}{5} \times \log_2 \frac{3}{5} = 0.971$$

$$S_{overcast} = [4+, 0-]$$

$$Entropy(S_{overcast}) = -\frac{4}{4} \times \log_2 \frac{4}{4} - \frac{0}{4} \times \log_2 \frac{0}{4} = 0$$

$$S_{rain} = [3+, 2-]$$

$$Entropy(S_{rain}) = -\frac{3}{5} \times \log_2 \frac{3}{5} - \frac{2}{5} \times \log_2 \frac{2}{5} = 0.971$$

$$Gain(S, Outlook) = Entropy(S) - \sum_{v \in \{sunny, overcast, rain\}} \frac{|S_v|}{|S|} \times Entropy(S_v)$$

$$Gain(S, Outlook) = 0.94 - \frac{5}{14} \times Entropy(S_{sunny}) - \frac{4}{14} \times Entropy(S_{overcast}) - \frac{5}{14} \times Entropy(S_{rain})$$

$$Gain(S, Outlook) = 0.94 - \frac{5}{14} \times 0.971 - \frac{4}{14} \times 0 - \frac{5}{14} \times 0.971 = 0.2464$$

Information Gain

$$\text{Entropy}(S) = 0.94$$

$$\text{Gain}(S, \text{Outlook}) = 0.2464$$

$$\text{Gain}(S, \text{Temp}) = 0.0289$$

Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rainy	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

Attribute: Temp
Values (Temp) = Hot, Mild, Cool

$$S_{Hot} = [2+, 2-]$$

$$\text{Entropy}(S_{Hot}) = 1$$

$$S_{Mild} = [4+, 2-]$$

$$\text{Entropy}(S_{Mild}) = -\frac{4}{6} \times \log_2 \frac{4}{6} - \frac{2}{6} \times \log_2 \frac{2}{6} = 0.9183$$

$$S_{Cool} = [3+, 1-]$$

$$\text{Entropy}(S_{Cool}) = -\frac{3}{4} \times \log_2 \frac{3}{4} - \frac{1}{4} \times \log_2 \frac{1}{4} = 0.8113$$

$$\text{Gain}(S, \text{Temp}) = \text{Entropy}(S) - \sum_{v \in \{\text{Hot}, \text{Mild}, \text{Cool}\}} \frac{|S_v|}{|S|} \times \text{Entropy}(S_v)$$

$$\text{Gain}(S, \text{Temp}) = 0.94 - \frac{4}{14} \times \text{Entropy}(S_{Hot}) - \frac{6}{14} \times \text{Entropy}(S_{Mild}) - \frac{4}{14} \times \text{Entropy}(S_{Cool})$$

$$\text{Gain}(S, \text{Temp}) = 0.94 - \frac{4}{14} \times 1 - \frac{6}{14} \times 0.9183 - \frac{4}{14} \times 0.8113 = 0.0289$$

Information Gain

Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rainy	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

Entropy (S) = 0.94

$$Gain(S, Outlook) = 0.2464$$

Gain (S,Temp) = 0.0289

$$Gain(S, Wind) = 0.048$$

Attribute: Wind
Values (Wind) = Weak, Strong

$$S_{Weak} = [6+, 2-]$$
$$\text{Entropy}(S_{Weak}) = -\frac{6}{8} \times \log_2 \frac{6}{8} - \frac{2}{8} \times \log_2 \frac{2}{8} = 0.811$$
$$S_{Strong} = [3+, 3-]$$
$$Entropy(S_{Strong}) = 1$$

$$Gain(S, Wind) = Entropy(S) - \sum_{v \in \{Weak, Strong\}} \frac{|S_v|}{|S|} \times Entropy(S_v)$$

$$Gain(S, Wind) = 0.94 - \frac{8}{14} \times Entropy(S_{Weak}) - \frac{6}{14} \times Entropy(S_{Strong})$$

$$Gain(S, Wind) = 0.94 - \frac{8}{14} \times 0.811 - \frac{6}{14} \times 1 = 0.048$$

Information Gain

Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rainy	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

$$Entropy(S) = 0.94$$

$$Gain(S, Outlook) = 0.2464$$

$$Gain(S, Temp) = 0.0289$$

$$Gain(S, Wind) = 0.048$$

$$Gain(S, Humidity) = 0.151$$

Attribute: Humidity
Values (Humidity) = High, Normal

$$S_{High} = [3+, 4-]$$

$$Entropy(S_{High}) = -\frac{3}{7} \times \log_2 \frac{3}{7} - \frac{4}{7} \times \log_2 \frac{4}{7} = 0.985$$

$$S_{Normal} = [6+, 1-]$$

$$Entropy(S_{Normal}) = -\frac{6}{7} \times \log_2 \frac{6}{7} - \frac{1}{7} \times \log_2 \frac{1}{7} = 0.592$$

$$Gain(S, Humidity) = Entropy(S) - \sum_{v \in \{High, Normal\}} \frac{|S_v|}{|S|} \times Entropy(S_v)$$

$$Gain(S, Humidity) = 0.94 - \frac{7}{14} \times Entropy(S_{High}) - \frac{7}{14} \times Entropy(S_{Normal})$$

$$Gain(S, Humidity) = 0.94 - \frac{7}{14} \times 0.985 - \frac{7}{14} \times 0.592 = 0.151$$

Information Gain

Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rainy	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

$$Entropy(S) = 0.94$$

$$Gain(S, Outlook) = 0.2464$$

$$Gain(S, Temp) = 0.0289$$

$$Gain(S, Wind) = 0.048$$

$$Gain(S, Humidity) = 0.151$$

Which attribute has the maximum predictability power with respect to the class label?

Assignment 1 : Task 2

- 1) **Text Preprocessing:** You are given two files. “train.csv” and “test.csv”. Both files have two columns “review” and “sentiment”. The first column is the input text while the second column is the label. You need to perform the following preprocessing steps on both files.
 - a. Read both csv files. You can use “pandas” or “polars” libraries to read it as dataframe in Python.
 - b. For each review, remove all
 tokens. You can use “re” library in Python that uses regular expressions or you can use any other library of your choice that allow string removal/replacement.
 - c. Change labels from text to integers (0 for negative and 1 for positive). You can do it directly in pandas dataframe or use scikit-learn label encoder.
 - d. Separate out labels and texts. You should have train texts, train labels, test texts, and test labels in separate variables.
 - e. For the preprocessed texts above, convert them to vectors based on frequency (counts) using scikit-learn. Do this for both train and test splits. Remember, you can only use train texts to determine the vocabulary of the dataset. **Hint:** You will have one vectorizer that has seen training data and will be used to transform the testing data.
- 2) **Feature Importance:**
 - a. For the preprocessed texts above, convert them to vectors based on binary occurrences of words using scikit-learn. Do this for both train and test splits. Save these vector representations in two variables.
 - b. Use information gain to calculate feature importance (words importance to be specific) for train split only. Sort the words based on the information gain values in descending order and output top 10 (most important) features (words). Output last 10 (least important) features (words). You can use **scikit-learn mutual_info_classif** method to calculate feature importance or write your own function to calculate information gain.

Book Reading

☐ Murphy – Chapter 1, Chapter 2.8