Intro to ML
Classification

Kjong Lehmann

Evaluating a classifier (Accuracy/Sensitivity/Specificity)

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}}$$

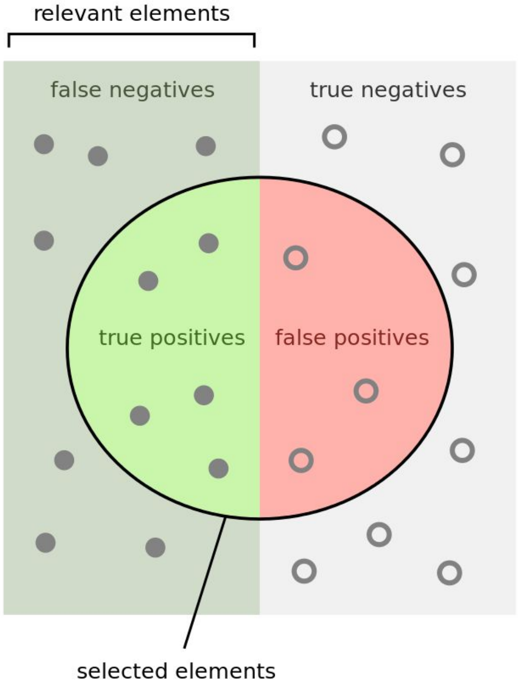
$$\text{accuracy} = \frac{\text{nr. correct predictions}}{\text{nr. total predictions}} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$$

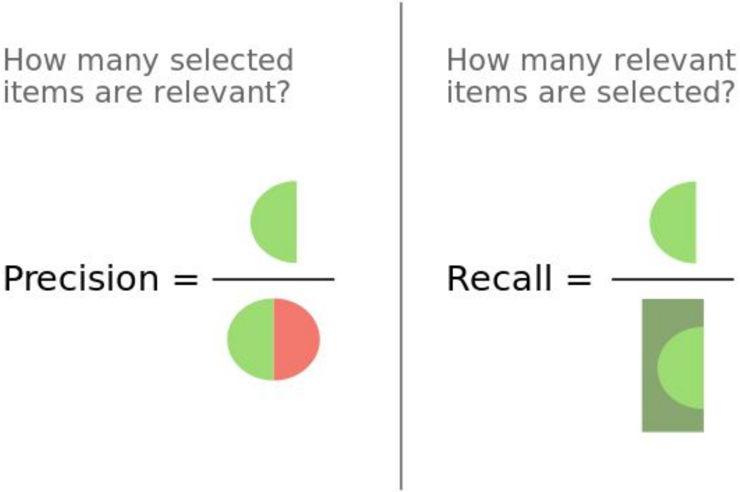
Evaluation Metrics (F1 score)

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

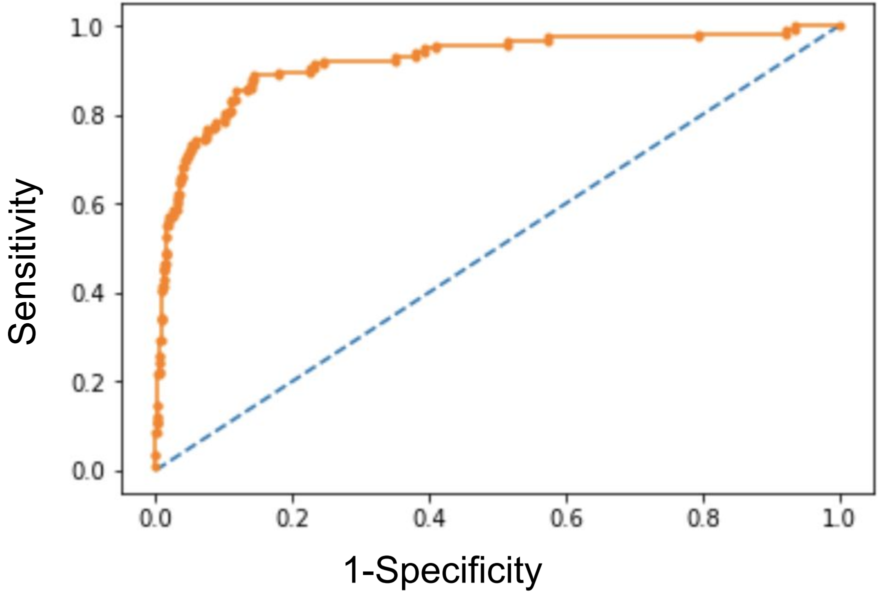
Evaluation Metrics (F1 score)





Source: wikipedia.org

ROC and AUC



Source: http://www.davidsbatista.net/blog/2018/08/19/NLP_Metrics/

Evaluation Metrics (F1 score)

$$\text{Precision} = \frac{tp}{tp + fp} \qquad \text{Recall} = \frac{tp}{tp + fn}$$

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Evaluation Metrics (F1 score Example)

Harmonic mean example : Doku-Cam

Evaluation Metrics (F1 score Example)

Harmonic mean example :

Average speed for the entire journey

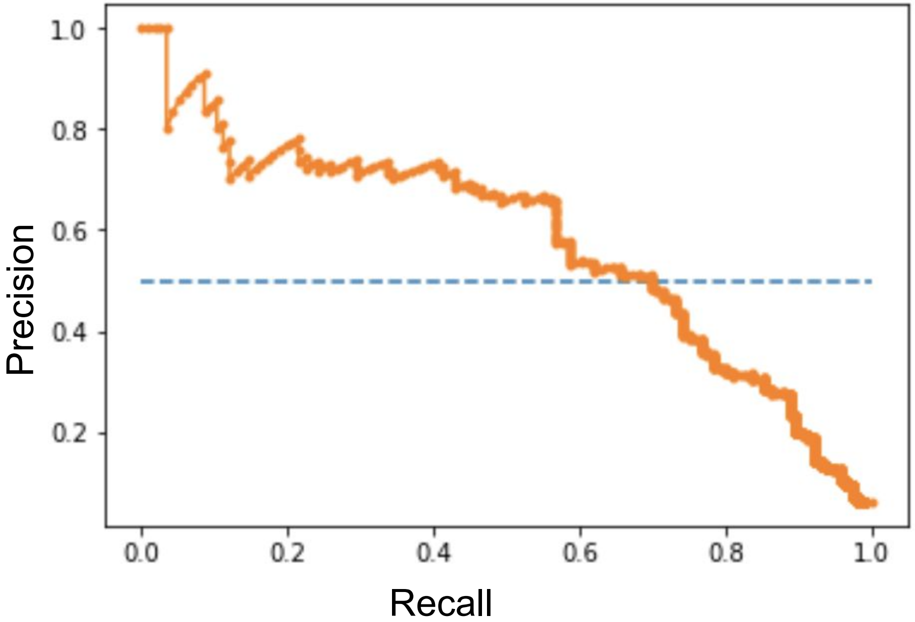
$$= \frac{\text{Total distance traveled}}{\text{Total time taken}}$$

$$= \frac{2d}{\frac{d}{x} + \frac{d}{y}} = \frac{2d}{\frac{yd + xd}{xy}} = \frac{2dxy}{d(x+y)}$$

$$= \frac{2xy}{x+y} \text{ (harmonic mean of } x \text{ and } y\text{)}$$

Source: wikipedia.org

PRC and AUPRC



What is the sense in PRAUC? Is this correct way to think about it?

Source: http://www.davidsbatista.net/blog/2018/08/19/NLP_Metrics/

Example

Would you recommend a classifier with 0.89 accuracy and 1.0 recall?

Example

Would you recommend a classifier with 0.89 accuracy and 1.0 recall?

No, since the high recall implies low precision

Imbalance issues example

$$\text{accuracy} = \frac{\text{nr. correct predictions}}{\text{nr. total predictions}} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$$

Assume 90% of data is positive?!?

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

Imbalanced Dataset

Small Demo

Loss Functions as surrogates of evaluation metric

- Classification Loss
- Perceptron Loss
- Hinge Loss

Evaluation of test and validation data
Loss function as surrogate to estimate w on training data

Loss Functions as surrogates of evaluation metric

Classification Losses

$L\text{perceptron}: \{-1, 1\} \times \mathbb{R} \rightarrow \mathbb{R}$ $\quad \mathbf{y}, f(\mathbf{x}) \rightarrow max(0, -\mathbf{y} f(\mathbf{x}))$

Find the best separation hyperplane

$L\text{hinge}: \{-1, 1\} \times \mathbb{R} \rightarrow \mathbb{R}$ $\quad \mathbf{y}, f(\mathbf{x}) \rightarrow max(0, 1 - \mathbf{y} f(\mathbf{x}))$

Find large separation margin

$L\text{logistic}: \{-1, 1\} \times \mathbb{R} \rightarrow \mathbb{R}$ $\quad \mathbf{y}, f(\mathbf{x}) \rightarrow log(1 + exp(-\mathbf{y} f(\mathbf{x})))$

Link to cross entropy and probabilistic interpretation, (cf. lecture logistic regression)

Which one more sensitive to outliers?

Break?

Encodings (Feature Scaling)

- SGD is scale sensitive
- Classifier relying on distances/similarities are scale sensitive
- Implications on test/train data

$$x' = \frac{x - \bar{x}}{\sigma}$$

Encodings (One-hot encoding)

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

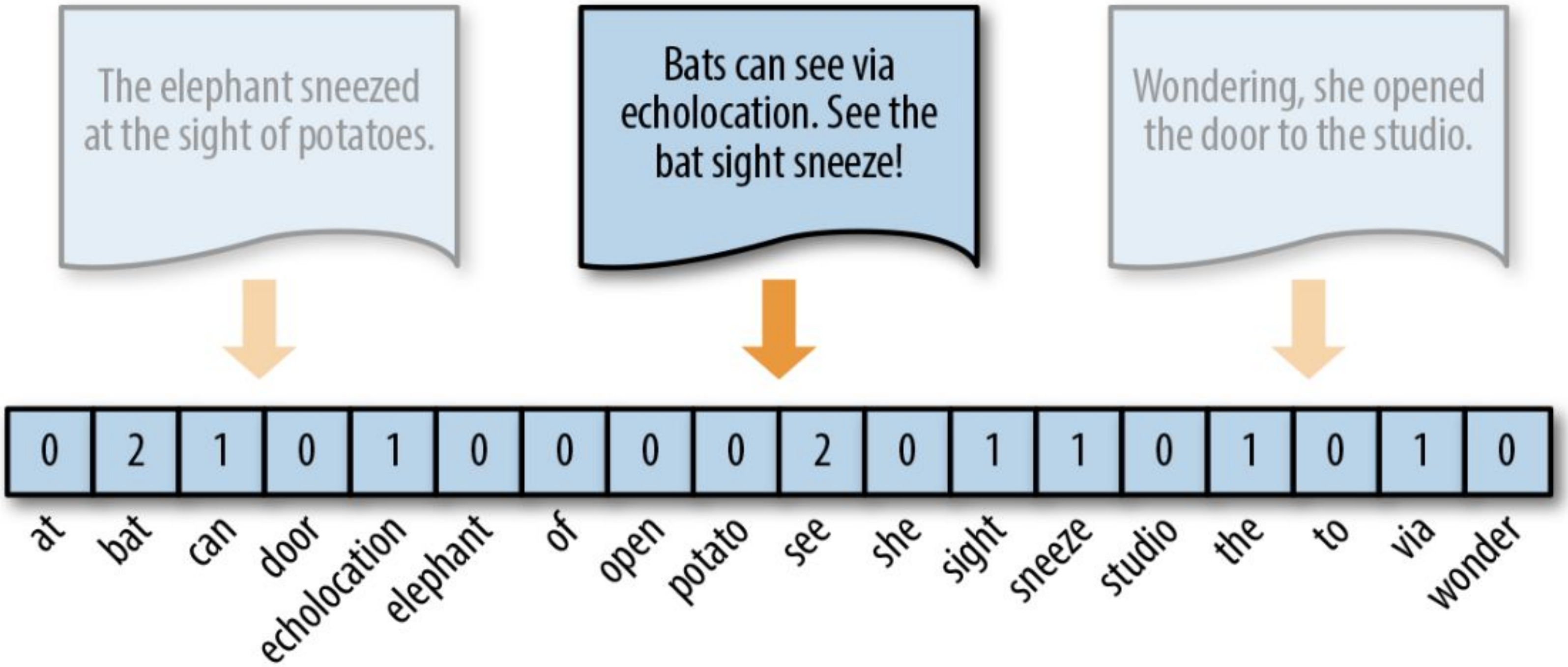
One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

Useful when utilizing some distance/similarity measure
Generalization?!?

Source: medium.com

Encodings (Bag of words)



Source: towradsdatascience.com

Feature Selection

- Important features can provide insights
- High model complexity can lead to overfitting (example)
- More features -> longer training time| memory (Trade-off)

Feature Selection (Univariate selection)

- F-Test or other test statistics, mutual information conditioned on label
  - Careful about linearity vs non-linearity

Feature Selection

- Greedily add or remove features from model
  - Cross-validation, importance/coefficient measure
- L1 penalty

Informative to understand algorithm?

End of Presentation
Start of Q&A