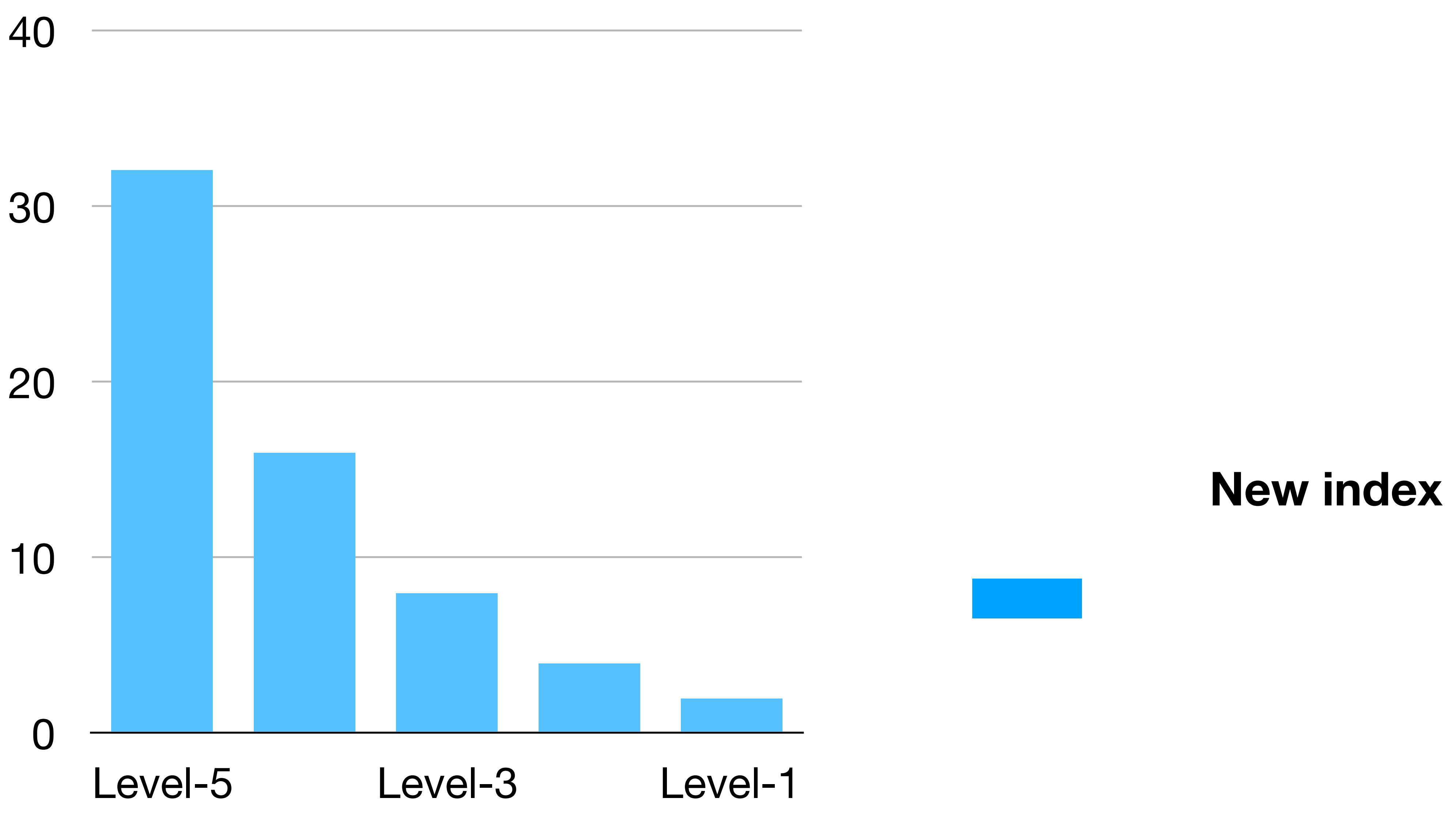
COMP90042
Web search and text analysis

Workshop Week 5

xudong.han@unimelb.edu.au
https://github.com/HanXudong/COMP90042_Workshops

Review

- Logarithmic index layout I/Os



This workshop

- Text classification
- N-gram language model
- Back-off and interpolation

What is text classification? Give some examples.

- Topic classification
- Sentiment analysis
- Authorship attribution
- Native-language identification
- Automatic fact-checking

Why is text classification generally a difficult problem? What are some hurdles that need to be overcome?

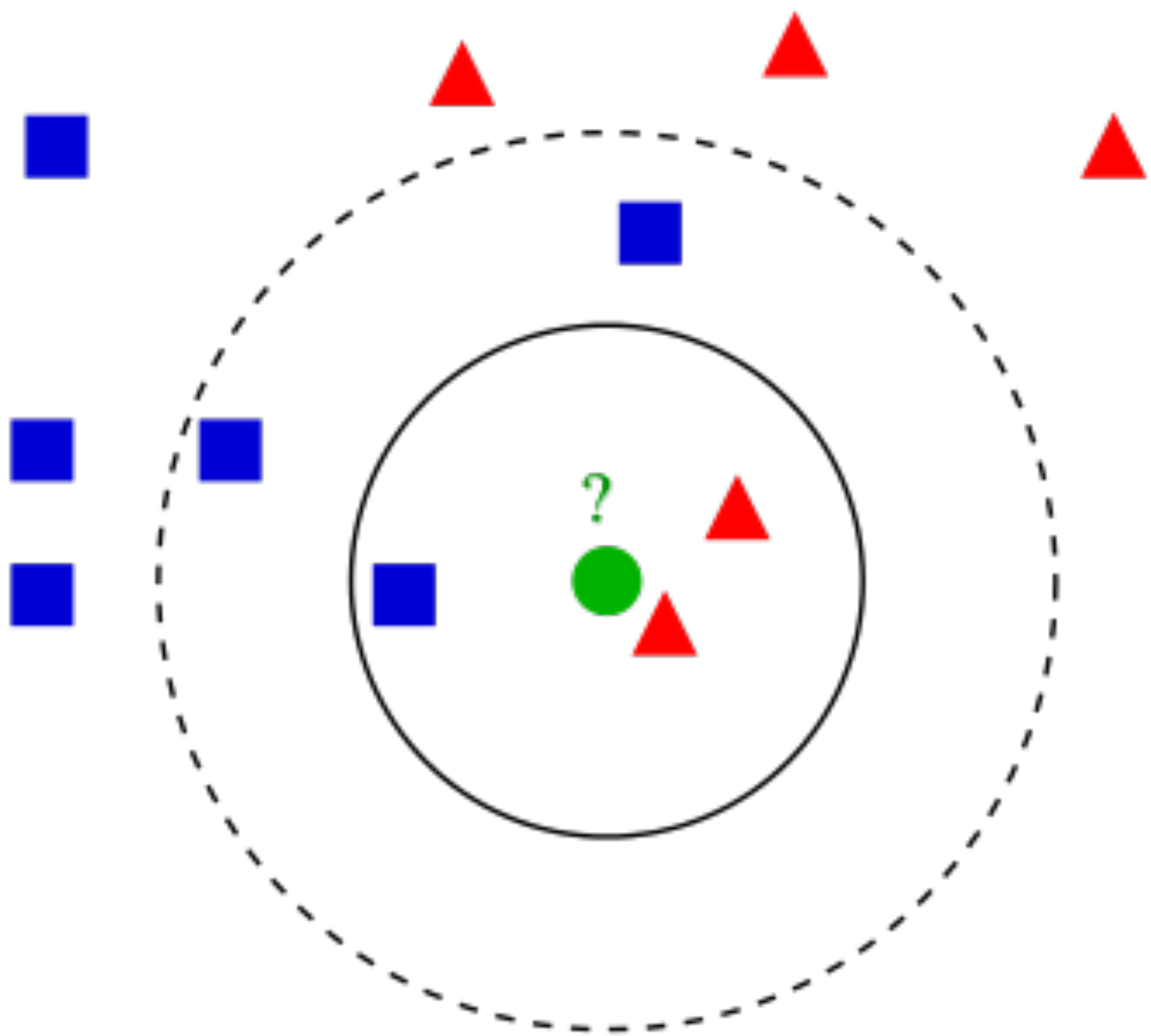
1. Identify a task of interest
2. Collect an appropriate corpus
3. Carry out annotation
4. Select features
5. Choose a machine learning algorithm
6. Tune hyper-parameters using held-out development data
7. Repeat earlier steps as needed
8. Train final model
9. Evaluate model on held-out test data

1-b **Consider some (supervised) text classification problem, and discuss whether the following (supervised) machine learning models would be suitable**

1. K-Nearest Neighbour using Euclidean distance
2. K-Nearest Neighbour using Cosine similarity
3. Decision Trees using information Gain
4. Naive Bayes
5. Logistic Regression
6. Support Vector Machine

KNN

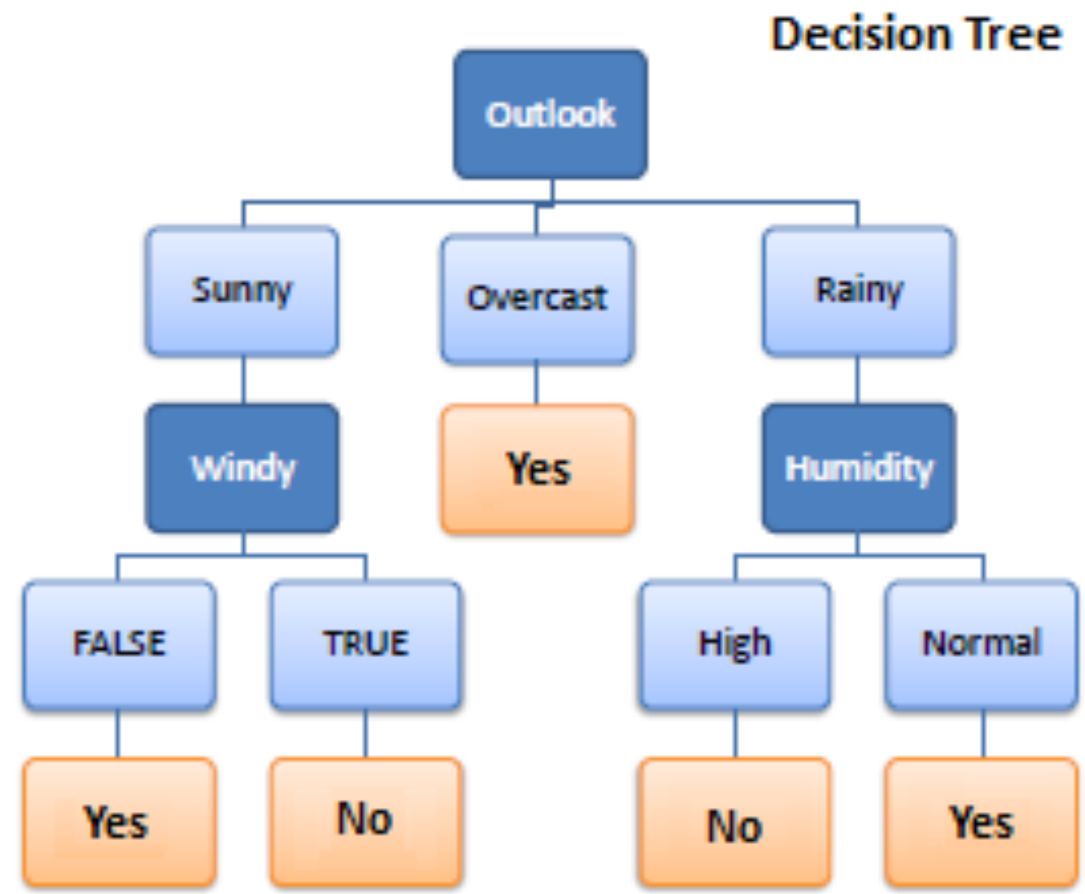
- Euclidean distance
- Cosine similarity



Decision tree

Predictors | Target

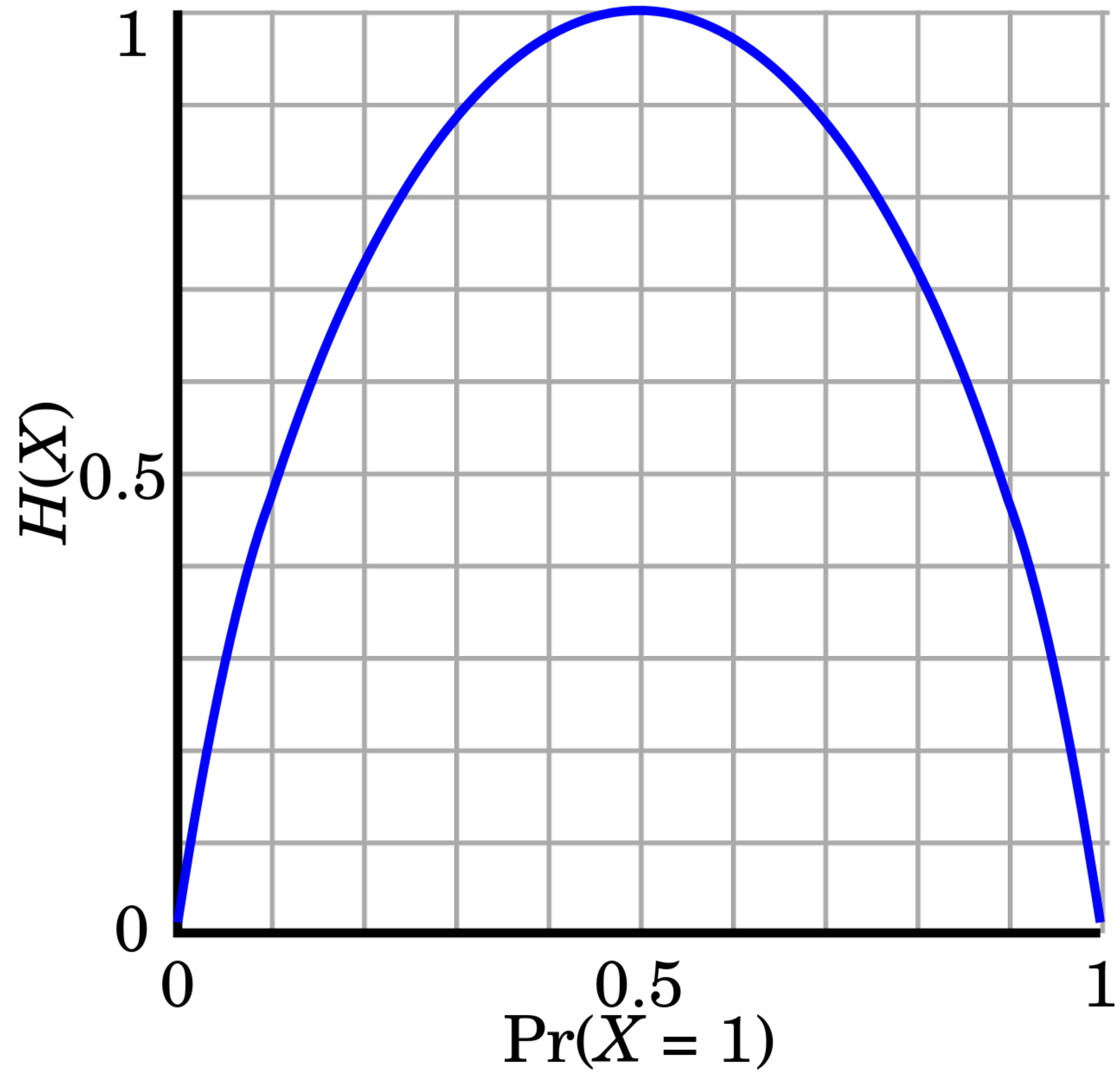
Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No



- Nodes correspond to features and leaves are final decision.
- The feature set is very large, and we might find spurious correlations.
- Information Gain is a poor choice because it tends to prefer rare features.

Decision tree

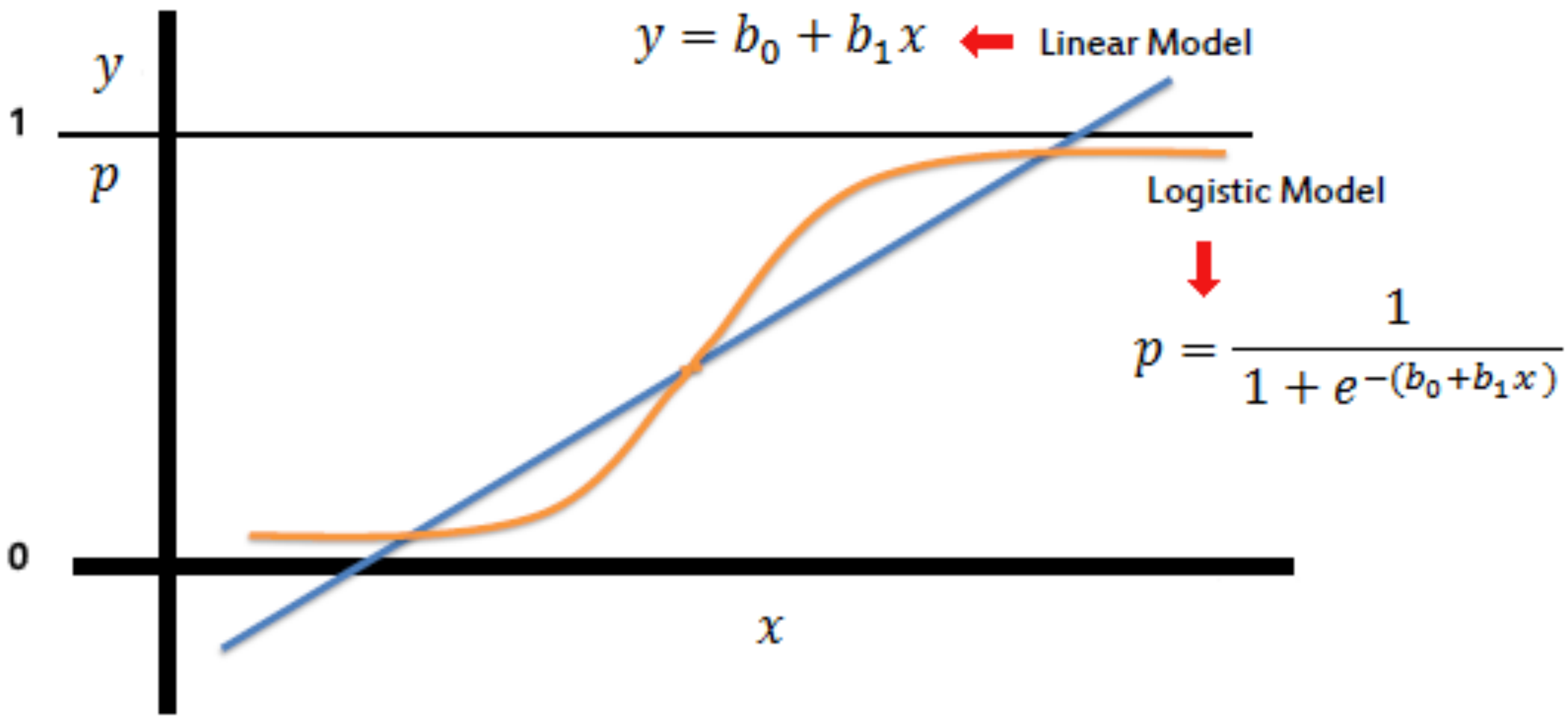
- https://www.saedsayad.com/decision_tree.htm



$$Entropy = -p\log_2 p - q\log_2 q$$

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

Logistic Regression



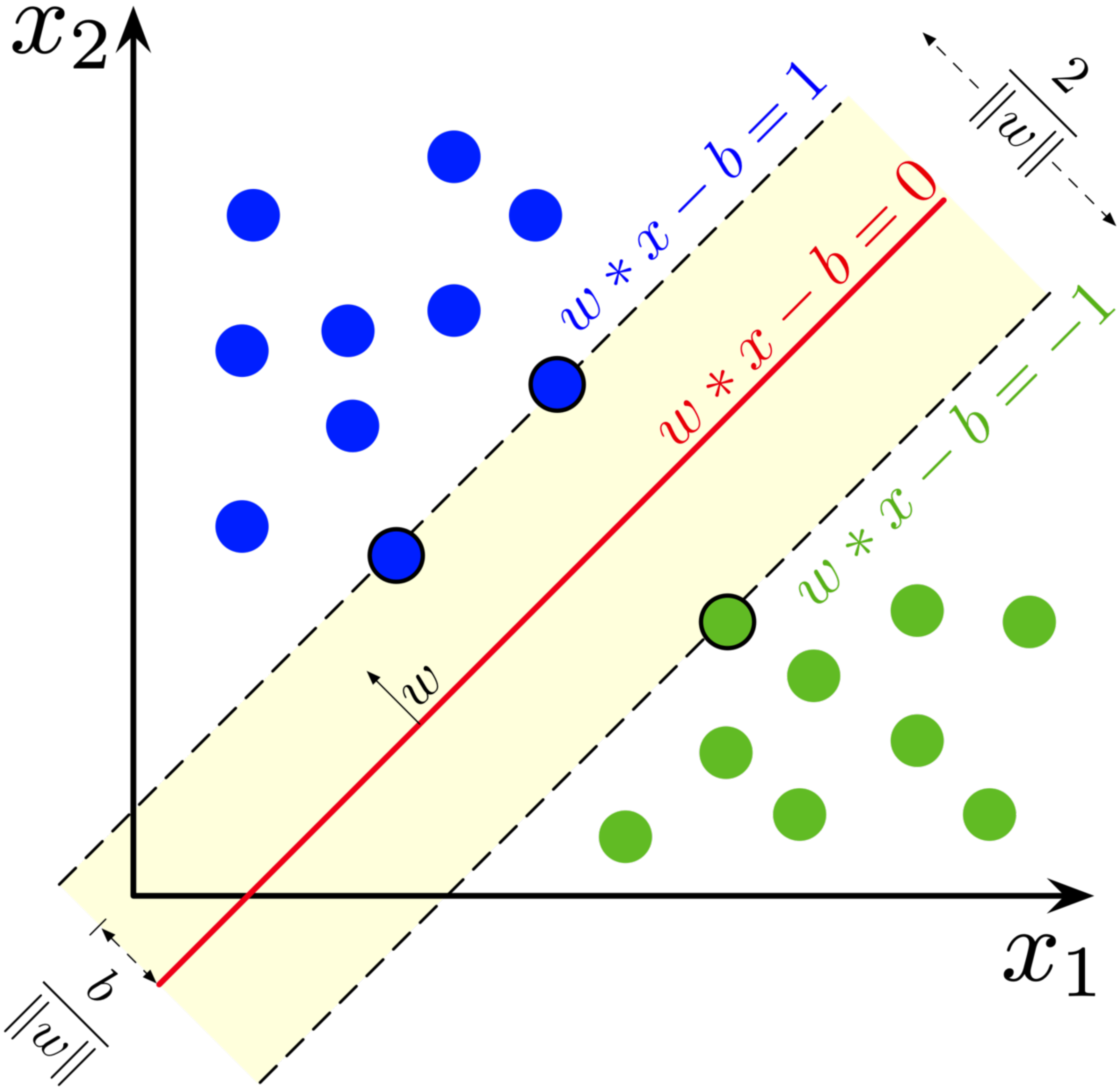
Naive Bayes

In machine learning, Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

$$Posterior \propto Prior \times Likelihood$$

$$P(\theta | x) = \frac{p(x | \theta)p(\theta)}{p(x)}$$

Support Vector Machines



maximum-margin classifier
-Hard margin
-Soft margin
-Kernel

Q2

2. For the following "corpus" of two documents:

   `1. how much wood would a wood chuck chuck if a wood chuck would chuck wood`
   `2. a wood chuck would chuck the wood he could chuck if a wood chuck would chuck wood`

   (a) Which of the following sentences: `a wood could chuck`; `wood would a chuck`; is more probable, accoding to:

      i. An unsmoothed uni-gram language model?
      ii. A uni-gram language model, with Laplacian ("add-one") smoothing?
      iii. An unsmoothed bi-gram language model?
      iv. A bi-gram language model, with Laplacian smoothing?
      v. An unsmoothed tri-gram language model?
      vi. A tri-gram language model, with Laplacian smoothing?

1. how much wood would a wood chuck chuck if a wood chuck would chuck wood

2. a wood chuck would chuck the wood he could chuck if a wood chuck would chuck wood

- <s> = sentence start; </s> = sentence end

- Q: Which of the following sentences:
  A: a wood could chuck;   B: wood would a chuck ;
  is more probable, according to: an uni-gram language model?

$$P(w_1, w_2, \ldots, w_m) = \prod_{i=1}^{m} P(w_i)$$

$$P(w_i) = \frac{C(w_i)}{\sum_{j=1}^{m} C(w_j)}$$

1. how much wood would a wood chuck chuck if a wood chuck would chuck wood

2. a wood chuck would chuck the wood he could chuck if a wood chuck would chuck wood

$$P(w_i) = \frac{C(w_i)}{\sum_{j=1}^{m} C(w_j)}$$

W	a	chuck	could	he	how	if	much	the	wood	would	</s>	Total
Count												
P												

1. how much wood would a wood chuck chuck if a wood chuck would chuck wood

2. a wood chuck would chuck the wood he could chuck if a wood chuck would chuck wood

$$P(w_i) = \frac{C(w_i)}{\sum_{j=1}^{V} C(w_j)}$$

W	a	chuck	could	he	how	if	much	the	wood	would	</s>	Total
Count	4	9	1	1	1	2	1	1	8	4	2	34
P	4/34	9/34	1/34	1/34	1/34	2/34	1/34	1/34	8/34	4/34	2/34	1

W	a	chuck	could	he	how	if	much	the	wood	would	</s>	Total
Count	4	9	1	1	1	2	1	1	8	4	2	34
P	4/34	9/34	1/34	1/34	1/34	2/34	1/34	1/34	8/34	4/34	2/34	1

$$P(w_1, w_2, \ldots, w_m) = \prod_{i=1}^{m} P(w_i)$$

- Q: Which of the following sentences:
  A: a wood could chuck; B: wood would a chuck ;
  is more probable, according to: an uni-gram language model?

$$P(A) = P(a)P(wood)P(could)P(chuck)P(</s>)$$
$$P(B) = P(wood)P(would)P(a)P(chuck)P(</s>)$$

What if Count(a)==0?

W	a	chuck	could	he	how	if	much	the	wood	would	</s>	Total
Count	4	9	1	1	1	2	1	1	8	4	2	34
P	5/45	10/45	2/45	2/45	2/45	3/45	2/45	2/45	9/45	5/45	3/45	1

$$P(w_1, w_2, \ldots, w_m) = \prod_{i=1}^{m} P(w_i) \qquad P_{add1}(w_i) = \frac{C(w_i) + 1}{\sum_{j=1}^{V} [C(w_j) + 1]}$$

- Q: Which of the following sentences:
  A: a wood could chuck; B: wood would a chuck ;
  is more probable, according to: an uni-gram language model with add-one smoothing?

$$P(A) = P(a)P(wood)P(could)P(chuck)P(\ </s>\ )$$

$$P(B) = P(wood)P(would)P(a)P(chuck)P(\ </s>\ )$$

1. how much wood would a wood chuck chuck if a wood chuck would chuck wood

2. a wood chuck would chuck the wood he could chuck if a wood chuck would chuck wood

- <s> = sentence start; </s> = sentence end

- Q: Which of the following sentences:
  A: a wood could chuck; B: wood would a chuck ;
  is more probable, according to: an bi-gram language model?

$$P(w_1, w_2, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_{i-1})$$

$$P(w_i \mid w_{i-1}) = \frac{C(w_{i-1} w_i)}{C(w_{i-1})}$$

1. how much wood would a wood chuck chuck if a wood chuck would chuck wood

2. a wood chuck would chuck the wood he could chuck if a wood chuck would chuck wood

$$P(w_i \mid w_{i-1}) = \frac{C(w_{i-1} w_i)}{C(w_{i-1})}$$

W_1,W_2	<s>a	<s>wood	chuck</s>
Count(w_1,w_2)	1	0	0
Count(w_1)	2	8	9
P(w_2|w_1)	1/2	0	0

$$P(w_1, w_2, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_{i-1})$$

- A: a wood could chuck;
- B: wood would a chuck ;

1. <s> how much wood would a wood chuck chuck if a wood chuck would chuck wood </s>

2. <s> a wood chuck would chuck the wood he could chuck if a wood chuck would chuck wood </s>

- <s> = sentence start; </s> = sentence end

- Q: Which of the following sentences:
  A: a wood could chuck; B: wood would a chuck ;
  is more probable, according to: an bi-gram language model with add-one smoothing?

$$P(w_1, w_2, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_{i-1})$$

$$P_{add1}(w_i \mid w_{i-1}) = \frac{C(w_{i-1}\ w_i) + 1}{\sum_{j=1}^{V} [C(w_{i-1}\ w_j) + 1]} = \frac{C(w_{i-1}\ w_i) + 1}{C(w_{i-1}) + V}$$

1. <s> <s> how much wood would a wood chuck chuck if a wood chuck would chuck wood </s>

2. <s> <s> a wood chuck would chuck the wood he could chuck if a wood chuck would chuck wood </s>

- <s> = sentence start; </s> = sentence end

- Q: Which of the following sentences:
  A: a wood could chuck;   B: wood would a chuck ;
  is more probable, according to: an tri-gram language model with add-one smoothing?

$$P(w_1, w_2, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \,|\, w_{i-2}\ w_{i-1})$$

$$P_{add1}(w_i \,|\, w_{i-2}\ w_{i-1}) = \frac{C(w_{i-2}\ w_{i-1}\ w_i) + 1}{C(w_{i-2}\ w_{i-1}) + V}$$

Q3: What does back–off mean, in the context of smoothing a language model? What does interpolation refer to?

- The idea in a Backoff model is to build an Ngram model based on an (N-1) model
- https://en.wikipedia.org/wiki/Katz%27s_back-off_model
- Interpolation: instead of just backing off to the non-zero Ngram, it is possible to take into account all Ngrams.
- Estimate lambdas from held-out dataset.