AGENDA

1. **Introduction. Plan of Proof**.
2. **Experiment**: From *Data-preparation* to *Feature-selection*.
3. **Experiment**: Model Training & Results.
4. **Theoretical Proof.**

4A. Formalize the Wisdom

“Text data is often linearly separable.”

Linear-separability is property of vectors, and text is **NOT inherently** vectors.
→ **“Text-data, when vectorized, …”**

Not any type of vectorization, the word “**often” implies:** BoW-based.
→ **“Text-data, when vectorized with BoW, …”**

“Linearly separable” is not enough, should be “**linearly classifiable**” instead.
WHY? If only about linear-separability, then all hyperplanes are OK?
NO, what we need is hyperplane that generalizes!

4A. Formalize the Wisdom

What we need to prove:
"Text data, vectorized with BoW-based, is linearly classifiable."

THE PLAN

1. Proof of linear-separability.

2. Prove that SVM found a hyperplane that generalizes.

4B. Linear Separability?

Short Answer

Yes. Because text-data, with BoW, has high dimensions.

Long Answer: Consider all possible partitions of ***p*** points into 2 classes, we have **2^p partitions**. How many are perfectly separated by a hyperplane?

Denote ***C(p, N)*** the number of such partitions of ***p*** points in ***N*** dimensions.

$$C(p+1, N) = 2\sum_{i=0}^{N} \binom{p}{i}$$

4B. Linear Separability

<u>MEANING</u>: Higher dimension → higher probability of being linearly-separable.
dim = size - 1 → **always!**

That's for *PERFECT linear separation*, what if we "tolerate" outliers?
→ **Even higher probability at smaller dimensions**.

<u>Conlusion</u>:
Text datasets in BoW-based representations, with its thousands dimensions, have very high chance of being **linearly-separable at some tolerance**.

4C. Does SVM-hyperplane generalizes?

Short Answer: YES. Because true-error is bounded, and that upper-bound decreases as the sample size increases.

Long Answer: At confidence $1-\delta$:, true-error is bounded by:

$$R(h) \leq R_{\text{emp}}(h) + \sqrt{\frac{8d_{\text{vc}}(\ln\frac{2m}{d_{\text{vc}}}+1)+8\ln\frac{4}{\delta}}{m}}$$

1. ***d_vc*** is the VC-dimension of hypothesis-space ***H***, i.e: **the hyperplanes**, measuring its complexity.

2. ***m*** is the sample size.

4C. Does SVM-hyperplane generalizes?

$$R(h) \leq R_{\text{emp}}(h) + \sqrt{\frac{8d_{\text{vc}}(\ln \frac{2m}{d_{\text{vc}}} + 1) + 8 \ln \frac{4}{\delta}}{m}}$$

1. As sample size increases, numerator increases **SLOWER** than denominator

→ The bound is **tighter**!

2. The smaller ***d_vc***, the tighter the bound.

Normally, ***d_vc(*** hyperplane **)** = ***dim*** + ***1***. In the case of text, it is much smaller, by embed our **prior-knowledge** into it.

4C. Does SVM-hyperplane generalizes?

<u>**Prior of Text**</u>: The appearance-or-not, important-or-not, of each word strongly determine the overall-meaning of a text.

→ Text-data, when viewed at the perspective of BoW, should have wide margin between the two classes.

<u>**MEANING**</u>**:** This prior relates to the following theorem to lower ***d_vc***(*hyperplane*)

- If data-points contained in a ball of radius ***R*.**
- With a margin ***p*** between the 2 classes, then: $d_{\mathrm{vc}} \leq \left\lceil \frac{R^2}{\rho^2} \right\rceil$

In scikit-learn, TF-IDF vectors are normalized, leads to ***R^2*** **< dim / 2**
→ ***d_vc*** is much smaller than **dim + 1.**

4C. Does SVM-hyperplane generalizes?

CONCLUSION

1. With Text-data, ***d_vc*****(**hyperplane**)** is not small, making in-sample error close to true-error.

2. As size increases, they are being more of the same

→ **It generalizes!**

3. This, complement with our good empirical result, proves that this isn't mere coincidence. **DONE!**

Thank You!

[REF] *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*, 1998, **Thorsten Joachims**.