Robust Models in Information Retrieval

Nedim Lipka

Benno Stein

Bauhaus-Universität Weimar [www.webis.de]

Robust Models in Information Retrieval

Outline

- Introduction
- Bias and Variance
- Robust Models in IR
- Summary
- Excursus: Bias Types

Introduction

Introduction

Classification Task

Given:

- feature space X with feature vectors $\mathbf{x}$
- classification function (closed form unknown) $c : X \to Y$
- sample $S = \{(\mathbf{x}, y) \mid \mathbf{x} \in X, y = c(\mathbf{x})\}$

Introduction

Classification Task

Given:

- feature space X with feature vectors $\mathbf{x}$
- classification function (closed form unknown) $c : X \to Y$
- sample $S = \{(\mathbf{x}, y) \mid \mathbf{x} \in X, y = c(\mathbf{x})\}$

Searched:

- hypothesis $h \in H$ that minimizes $\underbrace{P(h(\mathbf{x}) \neq c(\mathbf{x}))}_{\textit{err}(h)}$, the generalization error.

Introduction

Classification Task

Given:

- feature space X with feature vectors $\mathbf{x}$
- classification function (closed form unknown) $c : X \to Y$
- sample $S = \{(\mathbf{x}, y) \mid \mathbf{x} \in X, y = c(\mathbf{x})\}$

Searched:

- hypothesis $h \in H$ that minimizes $\underbrace{P(h(\mathbf{x}) \neq c(\mathbf{x}))}_{\textit{err}(h)}$, the generalization error.

Measuring effectiveness of h:

- $\textit{err}_S(h) = \frac{1}{|S|} \sum_{\mathbf{x} \in S} \textit{loss}_{0/1}(h(\mathbf{x}), c(\mathbf{x}))$

  $\textit{err}_S(h)$ is called test error if S is not used for the construction of h.

- $\textit{err}(h^*) := \min_{h \in H} \textit{err}(h)$ defines lower bound for $\textit{err}(h)$ ➜ restriction bias.

Introduction

Classification Task

Given:

- set O of real-world objects o
- feature space X with feature vectors $\mathbf{x}$
- classification function (closed form unknown) $c : X \to Y$
- sample $S = \{(\mathbf{x}, y) \mid \mathbf{x} \in X, y = c(\mathbf{x})\}$

Searched:

- hypothesis $h \in H$ that minimizes $\underbrace{P(h(\mathbf{x}) \neq c(\mathbf{x}))}_{\textit{err}(h)}$, the generalization error.

Measuring effectiveness of h:

- $\textit{err}_S(h) = \frac{1}{|S|} \sum_{\mathbf{x} \in S} \textit{loss}_{0/1}(h(\mathbf{x}), c(\mathbf{x}))$

  $\textit{err}_S(h)$ is called test error if S is not used for the construction of h.

- $\textit{err}(h^*) := \min_{h \in H} \textit{err}(h)$ defines lower bound for $\textit{err}(h)$ ➜ restriction bias.

Introduction

Model Formation Task

The process (the function) α for deriving $\mathbf{x}$ from o is called *model formation*.

$$\alpha : O \rightarrow X$$

Introduction

Model Formation Task

The process (the function) α for deriving $\mathbf{x}$ from o is called *model formation*.

$$\alpha : O \to X$$

Choosing between different model formation functions $\alpha_1, \ldots, \alpha_m$

➜ choosing between different feature spaces $X_{\alpha_1}, \ldots, X_{\alpha_m}$

➜ choosing between different hypotheses spaces $H_{\alpha_1}, \ldots, H_{\alpha_m}$

Introduction

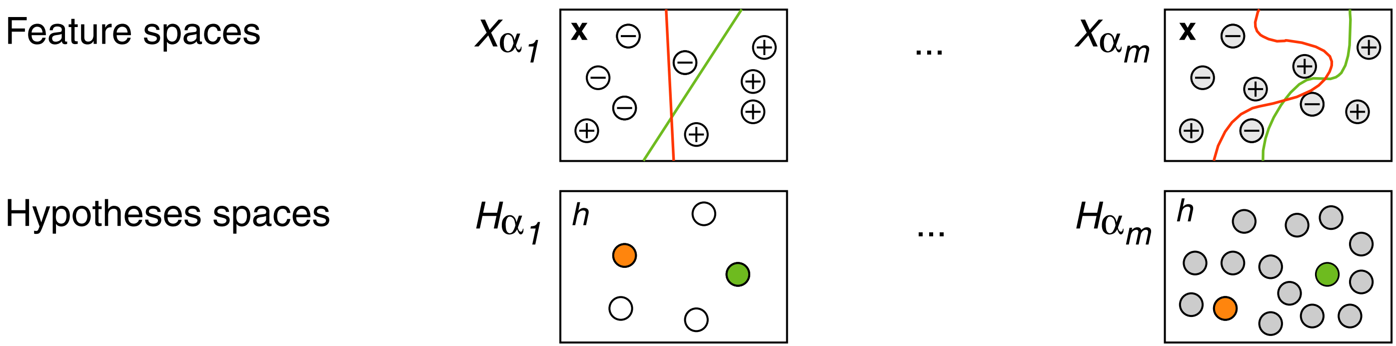
Model Formation Task

The process (the function) α for deriving $\mathbf{x}$ from o is called *model formation*.

$$\alpha : O \rightarrow X$$

Choosing between different model formation functions $\alpha_1, \ldots, \alpha_m$

- ➜ choosing between different feature spaces $X_{\alpha_1}, \ldots, X_{\alpha_m}$
  - ➜ choosing between different hypotheses spaces $H_{\alpha_1}, \ldots, H_{\alpha_m}$



Introduction

Model Formation Task

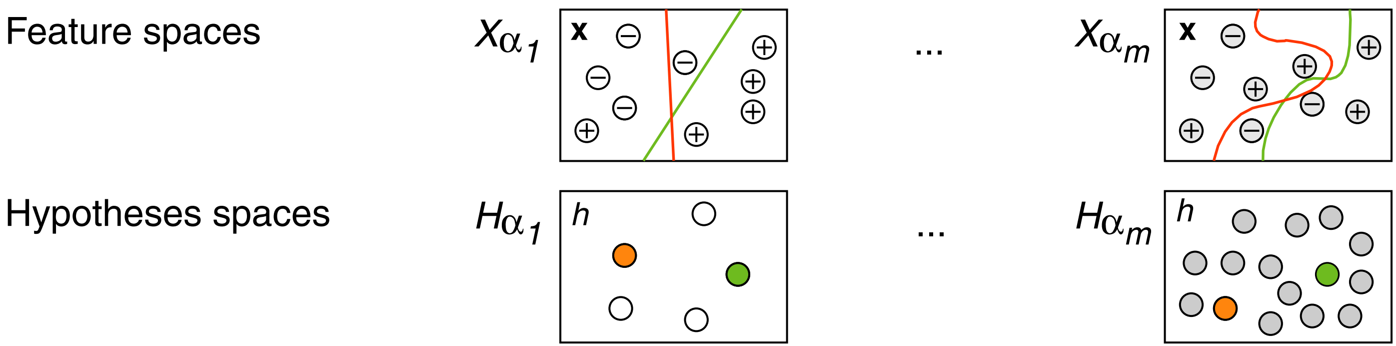
The process (the function) α for deriving $\mathbf{x}$ from o is called *model formation*.

$$\alpha : O \rightarrow X$$

Choosing between different model formation functions $\alpha_1, \ldots, \alpha_m$

➜ choosing between different feature spaces $X_{\alpha_1}, \ldots, X_{\alpha_m}$

➜ choosing between different hypotheses spaces $H_{\alpha_1}, \ldots, H_{\alpha_m}$

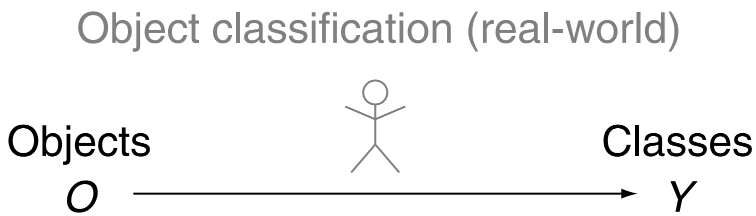


We call the model under α_1 being *more robust than* the model under α_2 $\Leftrightarrow$

$$\mathit{err}_S(h^*_{\alpha_1}) > \mathit{err}_S(h^*_{\alpha_2}) \quad \text{and} \quad \mathit{err}(h^*_{\alpha_1}) < \mathit{err}(h^*_{\alpha_2})$$

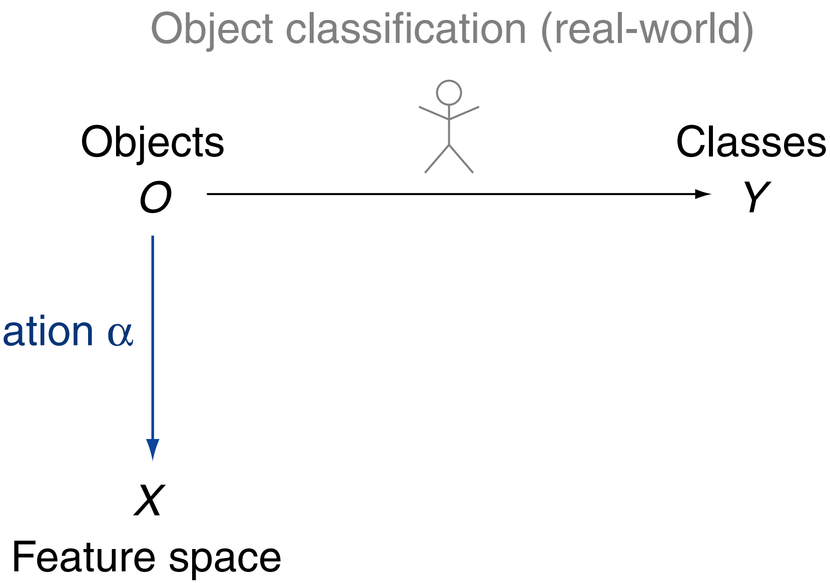
Introduction

The Whole Picture



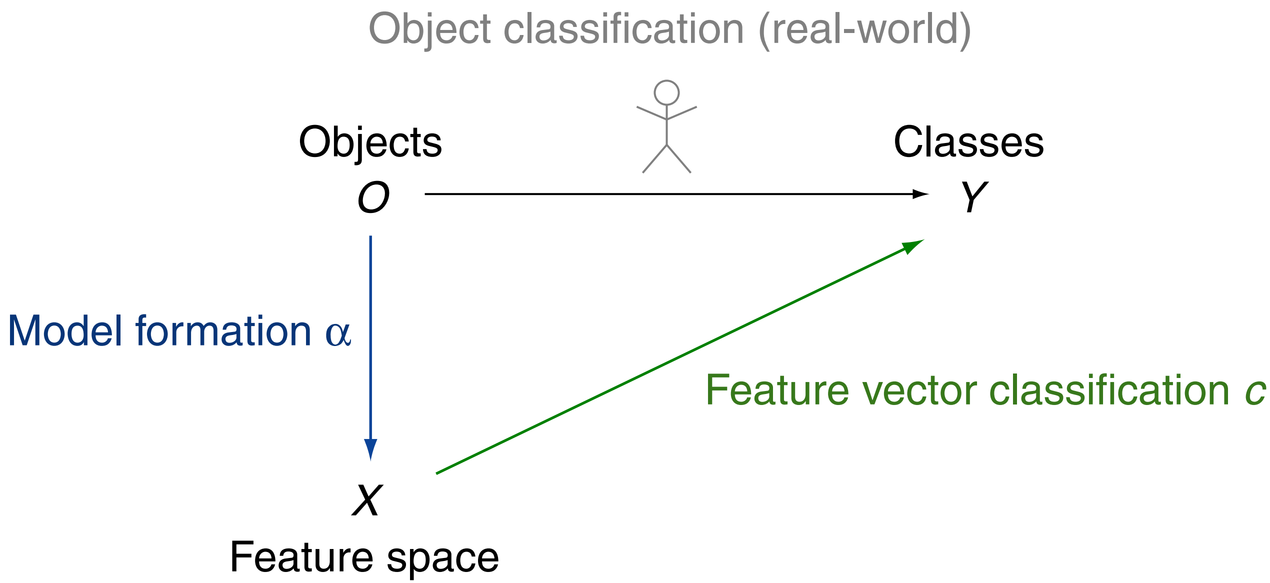
Introduction

The Whole Picture



Introduction

The Whole Picture



Learning means searching for a $h \in H$ such that $P(h(\mathbf{x}) \neq c(\mathbf{x}))$ is minimum.

Bias and Variance

Bias and Variance

Error Decomposition

Consider:

- A feature vector $\mathbf{x}$ and its predicted class label $\hat{y} = h(\mathbf{x})$, where
- h is characterized by a weight vector θ, where
- θ has been estimated based on a random sample $S = \{(\mathbf{x}, c(\mathbf{x})\}$.

➜ $\theta \equiv \theta(S)$, and hence $h \equiv h(\theta_S)$

Bias and Variance

Error Decomposition

Consider:

- A feature vector $\mathbf{x}$ and its predicted class label $\hat{y} = h(\mathbf{x})$, where
- h is characterized by a weight vector θ, where
- θ has been estimated based on a random sample $S = \{(\mathbf{x}, c(\mathbf{x})\}$.

➜ $\theta \equiv \theta(S)$, and hence $h \equiv h(\theta_S)$

Observations:

- A series of samples S_i, $S_i \subseteq U$, entails a series of hypotheses $h(\theta_i)$,
- giving for a feature vector $\mathbf{x}$ a series of class labels $\hat{y}_i = h(\theta_i, \mathbf{x})$.

➜ $\hat{y}$ is considered as a random variable, denoted as Z.

Bias and Variance

Error Decomposition

Consider:

- A feature vector $\mathbf{x}$ and its predicted class label $\hat{y} = h(\mathbf{x})$, where
- h is characterized by a weight vector θ, where
- θ has been estimated based on a random sample $S = \{(\mathbf{x}, c(\mathbf{x})\}$.

➜ $\theta \equiv \theta(S)$, and hence $h \equiv h(\theta_S)$

Observations:

- A series of samples S_i, $S_i \subseteq U$, entails a series of hypotheses $h(\theta_i)$,
- giving for a feature vector $\mathbf{x}$ a series of class labels $\hat{y}_i = h(\theta_i, \mathbf{x})$.

➜ $\hat{y}$ is considered as a random variable, denoted as Z.

Consequences:

- $\sigma^2(Z)$ is the variance of Z, (= variance of the prediction)
- $|\theta| : |S| \uparrow$ ➜ $\sigma^2(Z) \uparrow$
- $|S| : |U| \downarrow$ ➜ $\sigma^2(Z) \uparrow$

Bias and Variance

Error Decomposition (continued)

Let Z and Y denote the random variables for $\hat{y}$ ($= h(\theta_S, \mathbf{x})$) and y ($= c(\mathbf{x})$).

$$\mathit{MSE}(Z) = E((Z - Y)^2)$$

Bias and Variance

Error Decomposition (continued)

Let Z and Y denote the random variables for $\hat{y}$ $(= h(\theta_S, \mathbf{x}))$ and y $(= c(\mathbf{x}))$.

$$\begin{aligned} \textit{MSE}(Z) &= E((Z - Y)^2) \\ &= E(Z^2 - 2 \cdot Z \cdot Y + Y^2) \\ &= E(Z^2) \qquad\qquad - 2 \cdot E(Z \cdot Y) + E(Y^2) \end{aligned}$$

Bias and Variance

Error Decomposition (continued)

Let Z and Y denote the random variables for $\hat{y}$ ($= h(\theta_S, \mathbf{x})$) and y ($= c(\mathbf{x})$) .

$$
\begin{aligned}
\textit{MSE}(Z) &= E((Z-Y)^2) \\
&= E(Z^2 - 2 \cdot Z \cdot Y + Y^2) \\
&= E(Z^2) \qquad\qquad\qquad - 2 \cdot E(Z \cdot Y) + E(Y^2) \\
&= (E(Z))^2 + \sigma^2(Z) - 2 \cdot E(Z \cdot Y) + E(Y^2)
\end{aligned}
$$

Bias and Variance

Error Decomposition (continued)

Let Z and Y denote the random variables for $\hat{y}$ ($= h(\theta_S, \mathbf{x})$) and y ($= c(\mathbf{x})$).

$$
\begin{aligned}
\mathit{MSE}(Z) &= E((Z - Y)^2) \\
&= E(Z^2 - 2 \cdot Z \cdot Y + Y^2) \\
&= E(Z^2) \qquad\qquad\qquad - 2 \cdot E(Z \cdot Y) + E(Y^2) \\
&= (E(Z))^2 + \sigma^2(Z) - 2 \cdot E(Z \cdot Y) + E(Y^2) \\
&= (E(Z))^2 + \sigma^2(Z) - 2 \cdot E(Z \cdot Y) + (E(Y))^2 \quad + \sigma^2(Y)
\end{aligned}
$$

Bias and Variance

Error Decomposition (continued)

Let Z and Y denote the random variables for $\hat{y}$ $(= h(\theta_S, \mathbf{x}))$ and y $(= c(\mathbf{x}))$.

$$
\begin{aligned}
\mathit{MSE}(Z) &= E((Z-Y)^2) \\
&= E(Z^2 - 2\cdot Z\cdot Y + Y^2) \\
&= E(Z^2) \qquad\qquad\quad - 2\cdot E(Z\cdot Y) + E(Y^2) \\
&= (E(Z))^2 + \sigma^2(Z) - 2\cdot E(Z\cdot Y) + E(Y^2) \\
&= (E(Z))^2 + \sigma^2(Z) - 2\cdot E(Z\cdot Y) + (E(Y))^2 + \sigma^2(Y) \\
&= (E(Z))^2 \qquad\qquad\quad - 2\cdot E(Z\cdot Y) + (E(Y))^2 + \sigma^2(Y) + \sigma^2(Z) \\
&= (E(Z) - E(Y))^2 \qquad\qquad\qquad\qquad\qquad\quad + \sigma^2(Y) + \sigma^2(Z)
\end{aligned}
$$

Bias and Variance

Error Decomposition (continued)

Let Z and Y denote the random variables for $\hat{y}$ $(= h(\theta_S, \mathbf{x}))$ and y $(= c(\mathbf{x}))$.

$$
\begin{aligned}
\textit{MSE}(Z) &= E((Z-Y)^2) \\
&= E(Z^2 - 2 \cdot Z \cdot Y + Y^2) \\
&= E(Z^2) - 2 \cdot E(Z \cdot Y) + E(Y^2) \\
&= (E(Z))^2 + \sigma^2(Z) - 2 \cdot E(Z \cdot Y) + E(Y^2) \\
&= (E(Z))^2 + \sigma^2(Z) - 2 \cdot E(Z \cdot Y) + (E(Y))^2 + \sigma^2(Y) \\
&= (E(Z))^2 - 2 \cdot E(Z \cdot Y) + (E(Y))^2 + \sigma^2(Y) + \sigma^2(Z) \\
&= (E(Z) - E(Y))^2 + \sigma^2(Y) + \sigma^2(Z) \\
&= (E(Z-Y))^2 + \sigma^2(Z) + \sigma^2(Y) \\
&= (bias(Z))^2 + \sigma^2(Z) + \textit{IrreducibleError}
\end{aligned}
$$

Bias and Variance

Error Decomposition (continued)

Let Z and Y denote the random variables for $\hat{y}$ (= $h(\theta_S, \mathbf{x})$) and y (= $c(\mathbf{x})$) .

$$
\begin{aligned}
\mathit{MSE}(Z) &= E((Z-Y)^2) \\
&= E(Z^2 - 2 \cdot Z \cdot Y + Y^2) \\
&= E(Z^2) - 2 \cdot E(Z \cdot Y) + E(Y^2) \\
&= (E(Z))^2 + \sigma^2(Z) - 2 \cdot E(Z \cdot Y) + E(Y^2) \\
&= (E(Z))^2 + \sigma^2(Z) - 2 \cdot E(Z \cdot Y) + (E(Y))^2 + \sigma^2(Y) \\
&= (E(Z))^2 - 2 \cdot E(Z \cdot Y) + (E(Y))^2 + \sigma^2(Y) + \sigma^2(Z) \\
&= (E(Z) - E(Y))^2 + \sigma^2(Y) + \sigma^2(Z) \\
&= (E(Z-Y))^2 + \sigma^2(Z) + \sigma^2(Y) \\
&= (bias(Z))^2 + \sigma^2(Z) + \mathit{IrreducibleError}
\end{aligned}
$$

If Y is constant:

$$= (E(Z) - Y)^2 + \sigma^2(Z)$$

Bias and Variance

Error Decomposition (continued)

Let Z and Y denote the random variables for $\hat{y}$ ($= h(\theta_S, \mathbf{x})$) and y ($= c(\mathbf{x})$).

$$
\begin{aligned}
\textit{MSE}(Z) &= E((Z - Y)^2) \\
&= E(Z^2 - 2 \cdot Z \cdot Y + Y^2) \\
&= E(Z^2) - 2 \cdot E(Z \cdot Y) + E(Y^2) \\
&= (E(Z))^2 + \sigma^2(Z) - 2 \cdot E(Z \cdot Y) + E(Y^2) \\
&= (E(Z))^2 + \sigma^2(Z) - 2 \cdot E(Z \cdot Y) + (E(Y))^2 + \sigma^2(Y) \\
&= (E(Z))^2 - 2 \cdot E(Z \cdot Y) + (E(Y))^2 + \sigma^2(Y) + \sigma^2(Z) \\
&= (E(Z) - E(Y))^2 + \sigma^2(Y) + \sigma^2(Z) \\
&= (E(Z - Y))^2 + \sigma^2(Z) + \sigma^2(Y) \\
&= (bias(Z))^2 + \sigma^2(Z) + \textit{IrreducibleError}
\end{aligned}
$$

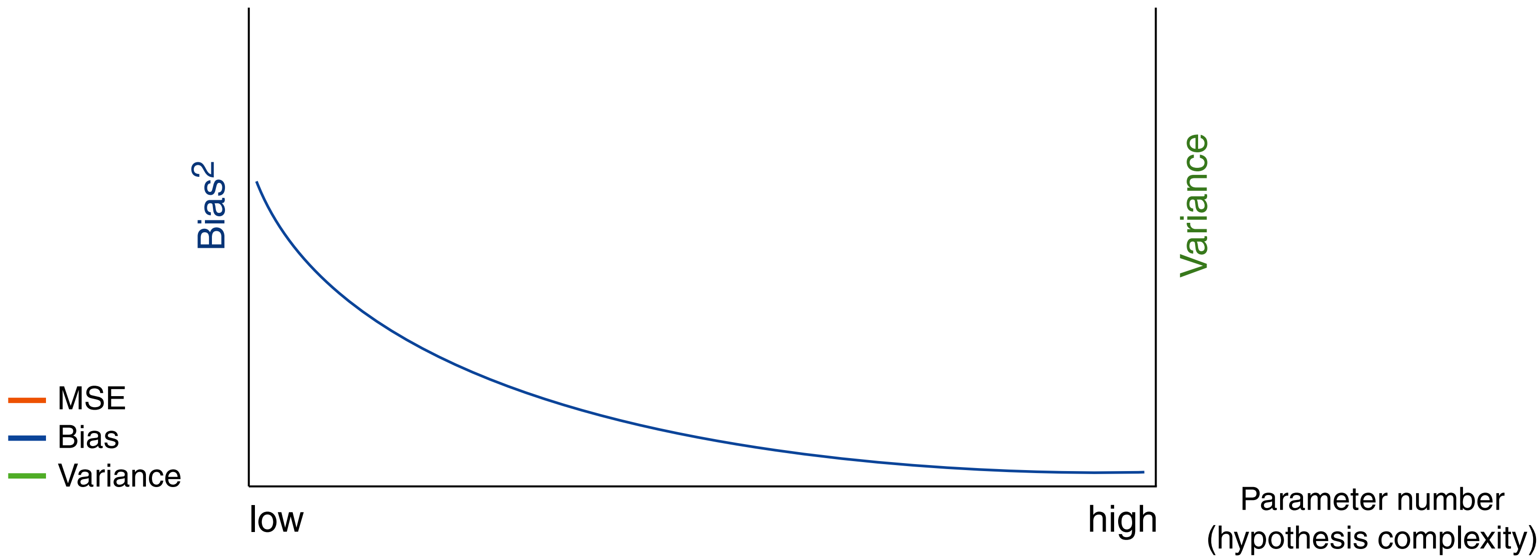
If Y is constant:

$$
= (E(Z) - Y)^2 + \sigma^2(Z)
$$

When analyzing *MSE*, $bias$, and σ^2 of a classifier h, the average over all examples of the test set is taken.

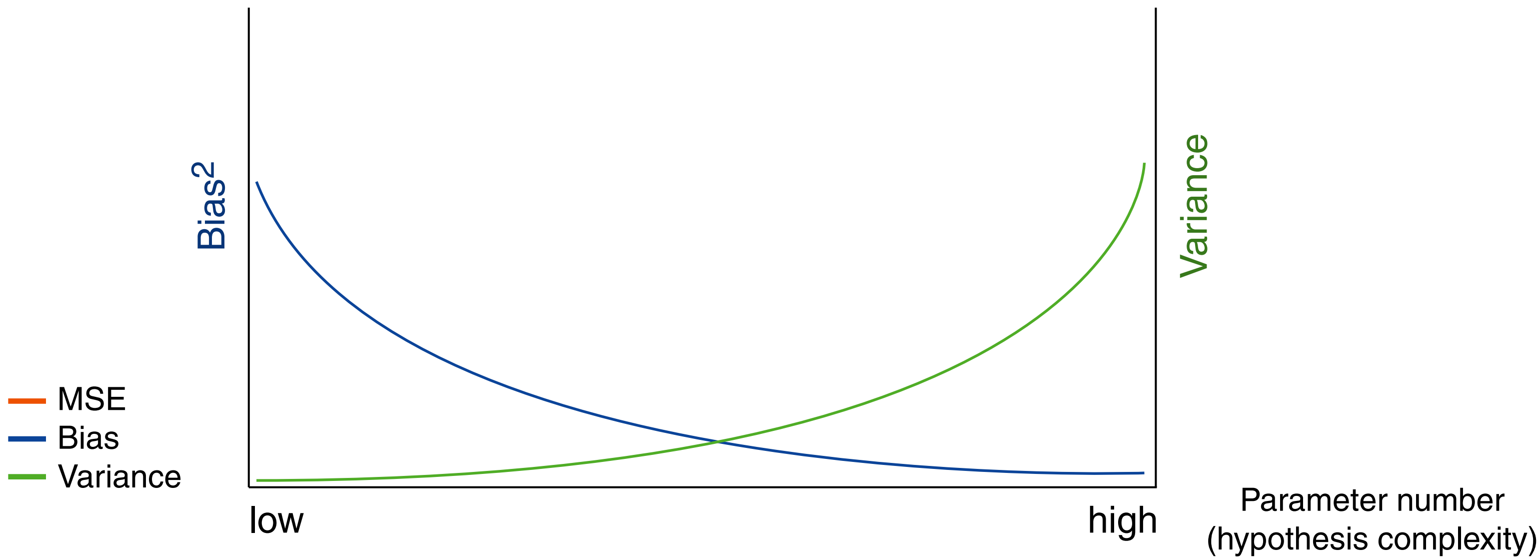
Bias and Variance

Illustration



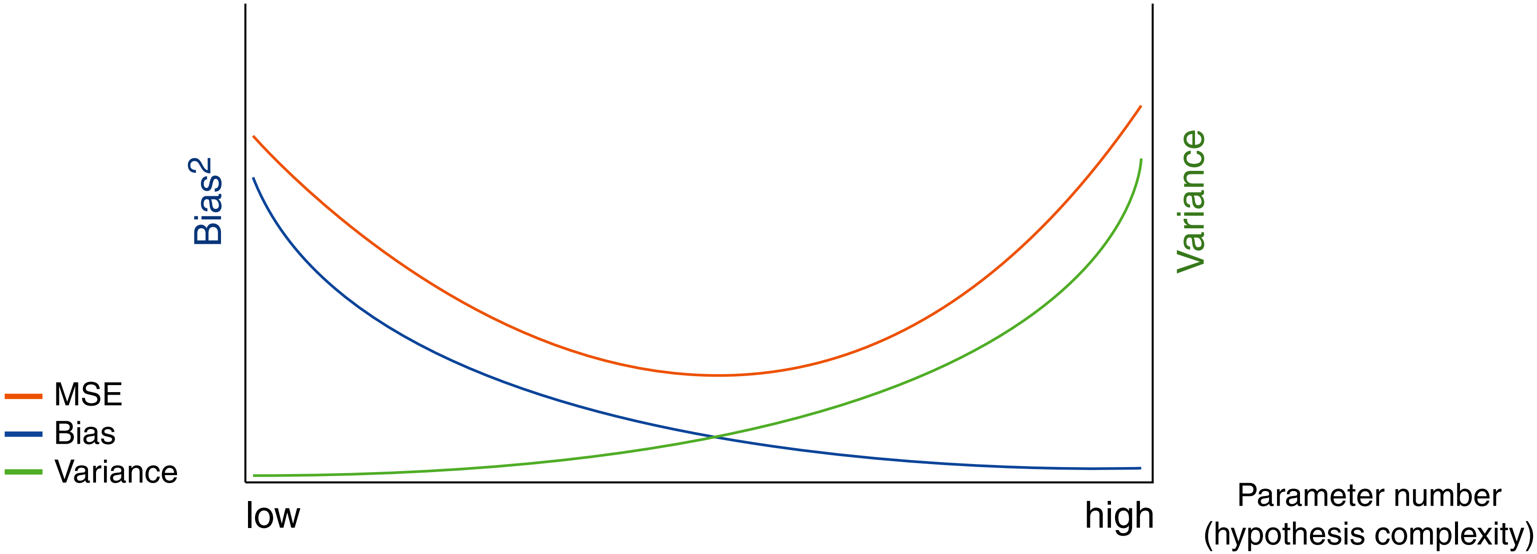
Bias and Variance

Illustration



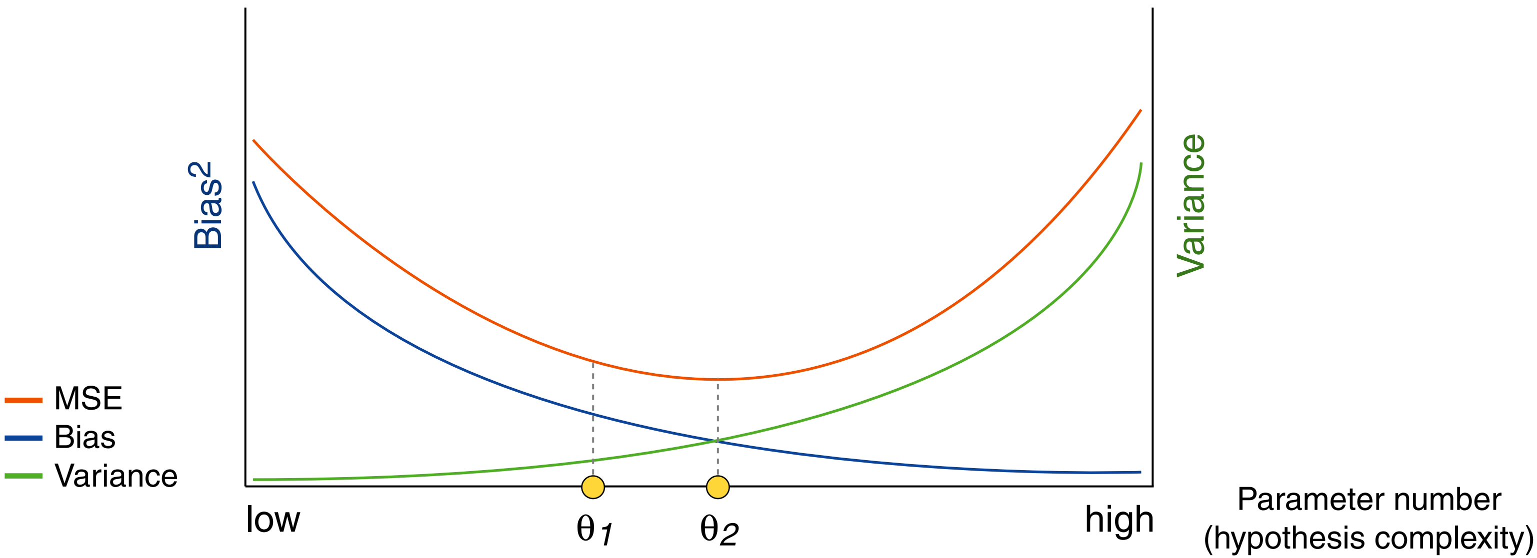
Bias and Variance

Illustration



Bias and Variance

Illustration

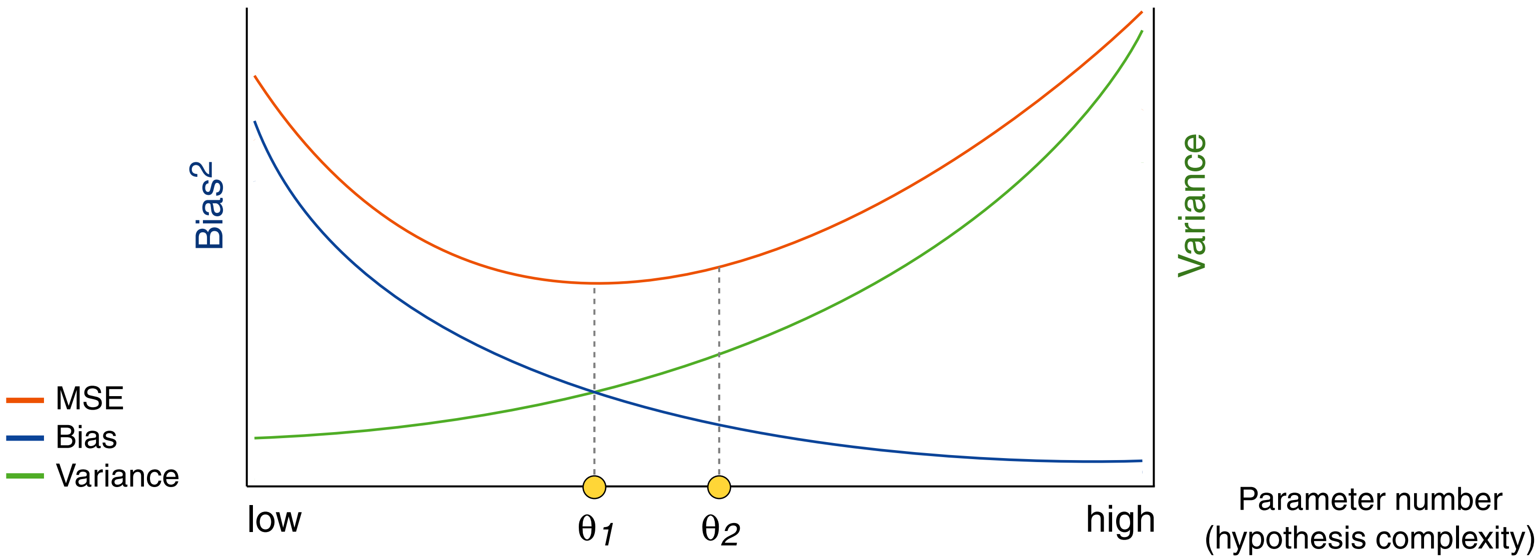


$$\textit{err}_S(h^*_{\alpha_1}) > \textit{err}_S(h^*_{\alpha_2})$$

Comparing two model-classifier-combinations under a sample S.

Bias and Variance

Illustration



$$\textit{err}(h^*_{\alpha_1}) < \textit{err}(h^*_{\alpha_2})$$

The same model-classifier-combinations under a sample S', with $|S'| \gg |S|$.

➜ The model under α_1 is more robust than the model under α_2.

Bias and Variance

Preliminary Summary

- Even when properly choosing training and test sets, a model selection decision may not be justified by error minimization.
- Rationale: the concept of representativeness gets lost for extreme ratios between the sample size and an application set in the wild.
  (consider working against the web)

➜ The bias of the less complex classifier is over-estimated.

➜ The variance of the more complex classifier is under-estimated.

- This behavior is consistent with the concept of the bias-variance-tradeoff.

Robust Models in IR

Robust Models in IR

Case Study I: Text Categorization

The model under α_1 is *more robust than* the model under α_2 $\Leftrightarrow$

$$\textit{err}_S(h^*_{\alpha_1}) > \textit{err}_S(h^*_{\alpha_2}) \qquad \text{and} \qquad \textit{err}(h^*_{\alpha_1}) < \textit{err}(h^*_{\alpha_2})$$

Experiment rationale:

- Topic classification for the web is learned on extremely small samples.
- The web generalization error of a classifier h cannot be computed.

➜ *err*(h) is usually unknown.

➜ Study the effect with a large (test) corpus in the role of the web by comparing *err*$_S(h_\alpha)$ and *err*(h_α) for different α.

Robust Models in IR

Case Study I: Text Categorization

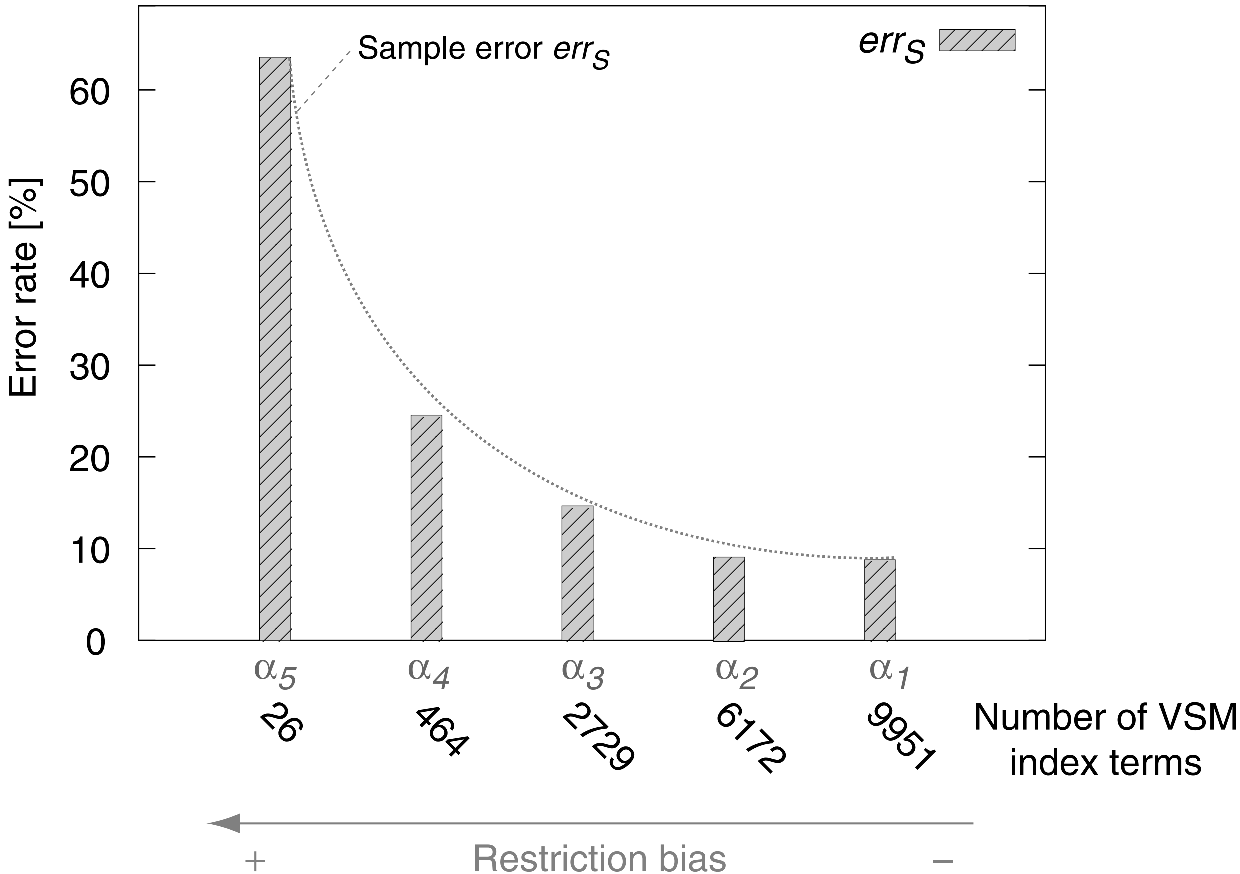
Experiment setup 1:

❑ Corpus	RCV1
❑ Corpus Size	663 768 documents
❑ Considered classes	corporate (292 348), economics (51 148), government (161 523), market (158 749)
❑ Sample size	800, drawn i.i.d. from RCV1
❑ Ratio sample and corpus	0.0012
❑ Inductive learner	SVM with linear kernel
❑ Model formation functions α	5 VSM variants

1. α_1: $V = \{[a\text{-}z]^5 *\}$, $|V| = 9951$
2. α_2: $V = \{[a\text{-}z]^4 *\}$, $|V| = 6172$
3. α_3: $V = \{[a\text{-}z]^3 *\}$, $|V| = 2729$
4. α_4: $V = \{[a\text{-}z]^2 *\}$, $|V| = 464$
5. α_5: $V = \{[a\text{-}z] *\}$, $|V| = 26$

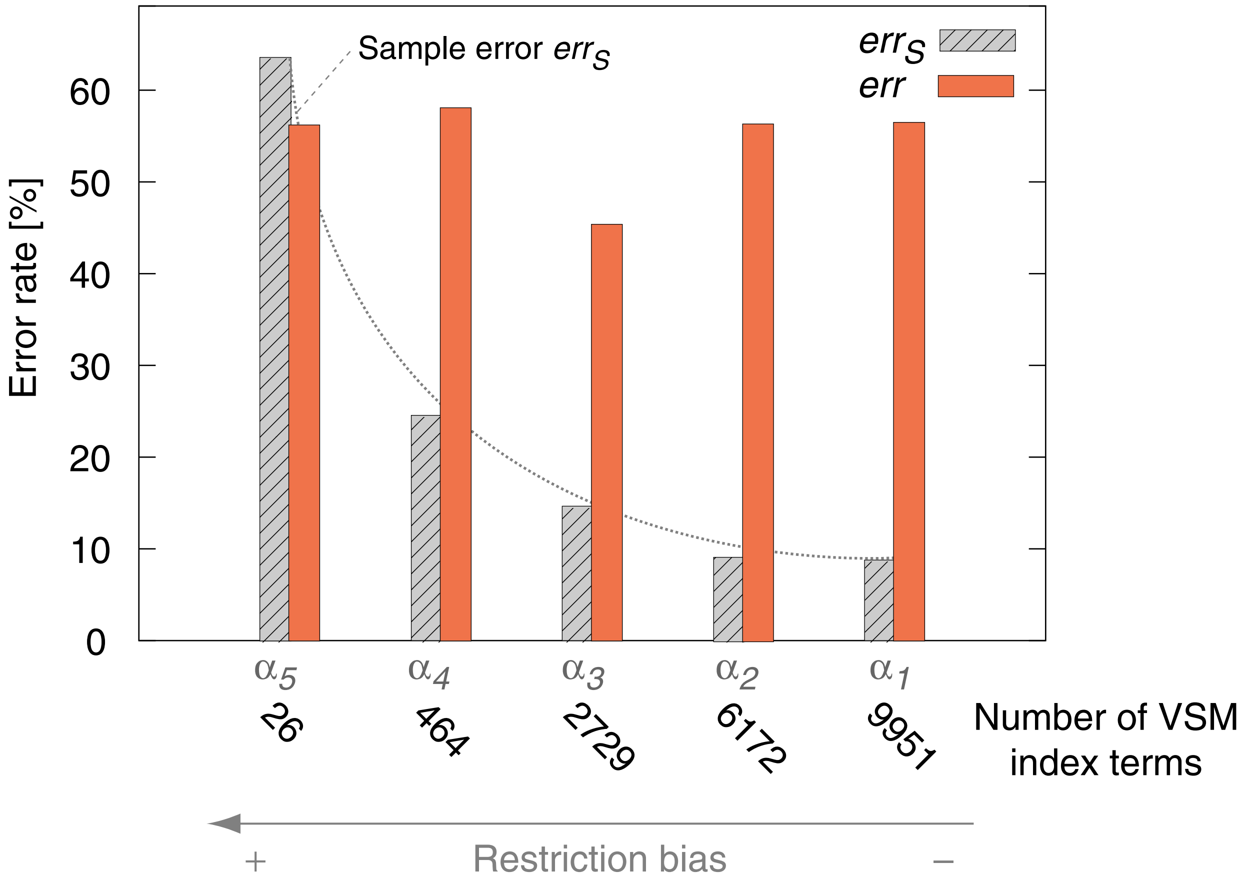
Robust Models in IR

Case Study I: Text Categorization



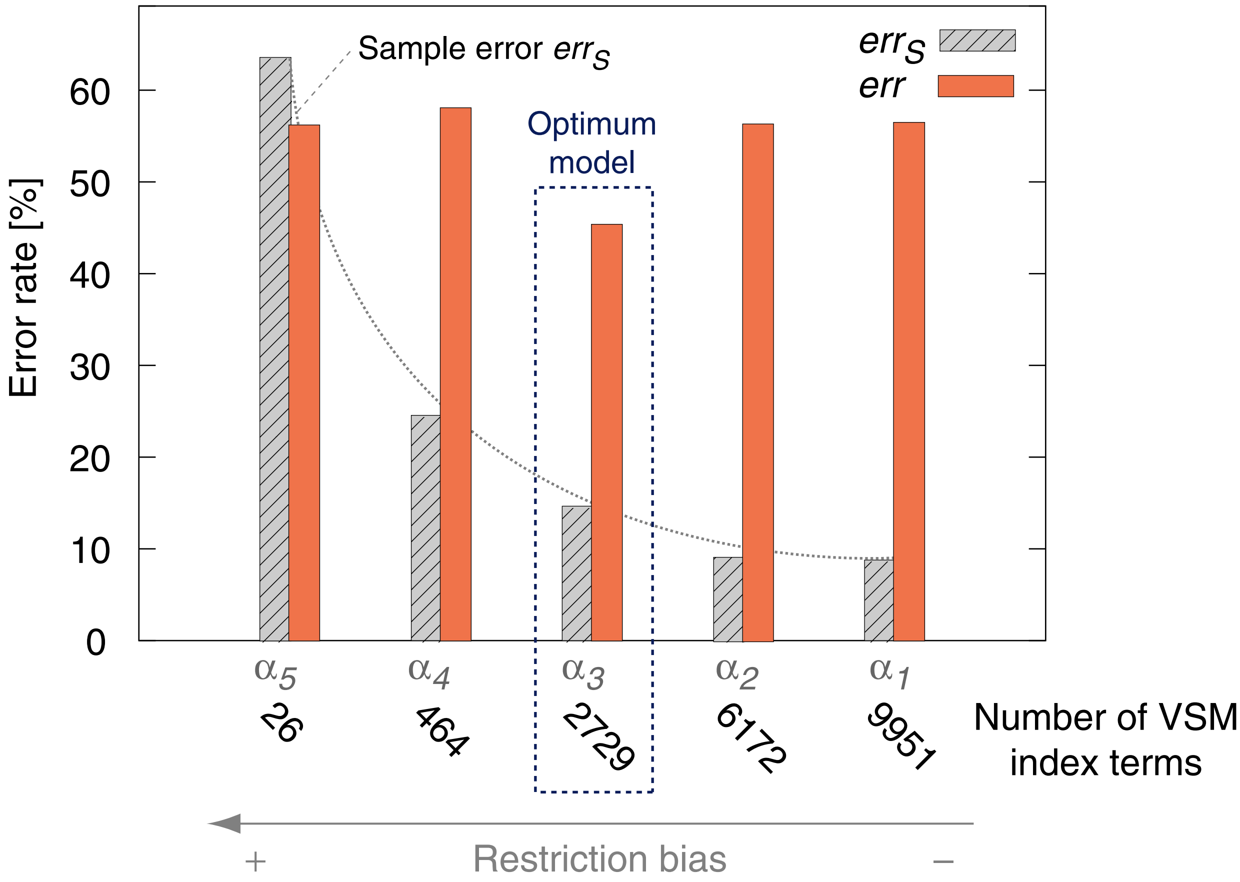
Robust Models in IR

Case Study I: Text Categorization



Robust Models in IR

Case Study I: Text Categorization



Robust Models in IR

Case Study I: Text Categorization

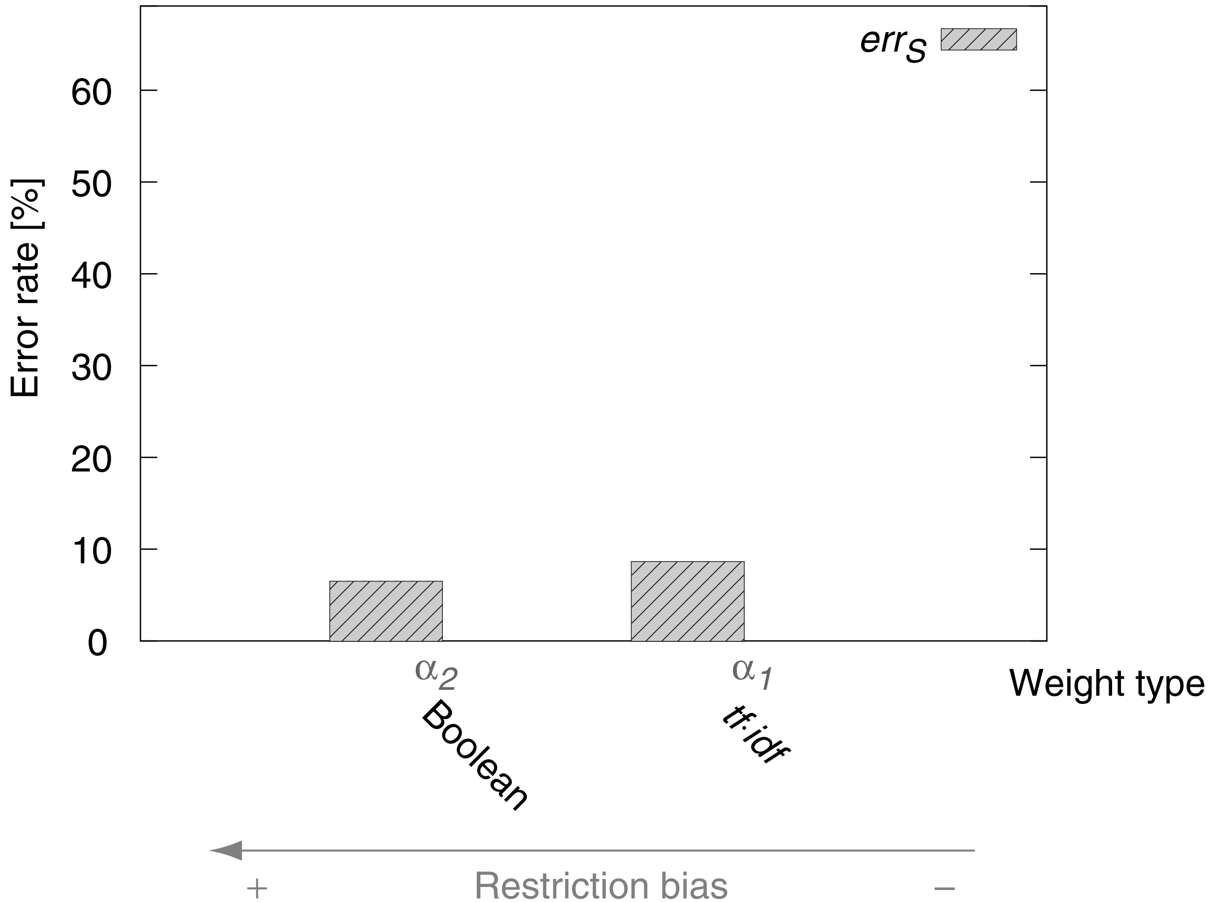
Experiment setup 2:

❑ Corpus	RCV1
❑ Corpus Size	663 768 documents
❑ Considered classes	corporate (292 348), economics (51 148), government (161 523), market (158 749)
❑ Sample size	800, drawn i.i.d. from RCV1
❑ Ratio sample and corpus	0.0012
❑ Inductive learner	SVM with linear kernel
❑ Model formation functions α	2 VSM variants

1. α_1: *tf·idf* weighting scheme
2. α_2: Boolean weighting scheme

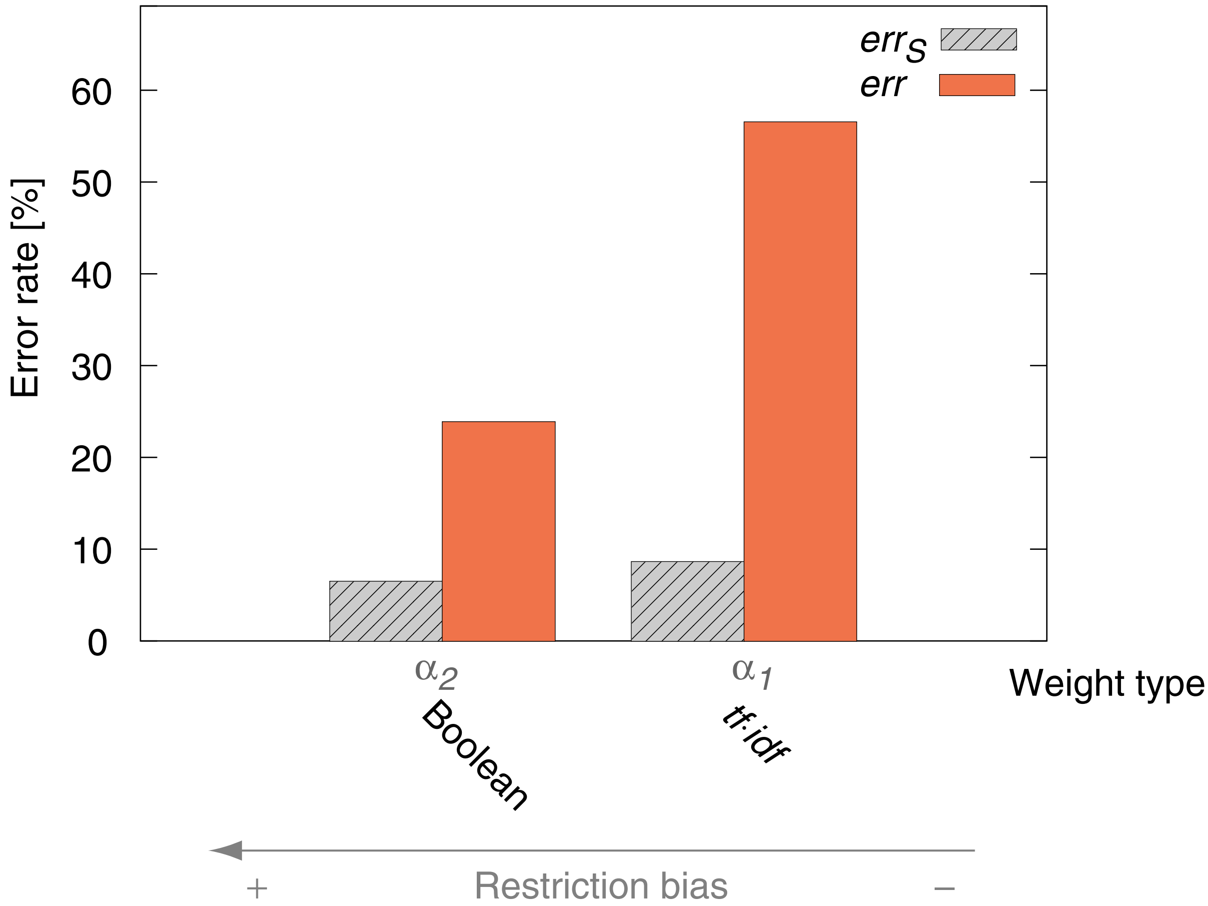
Robust Models in IR

Case Study I: Text Categorization



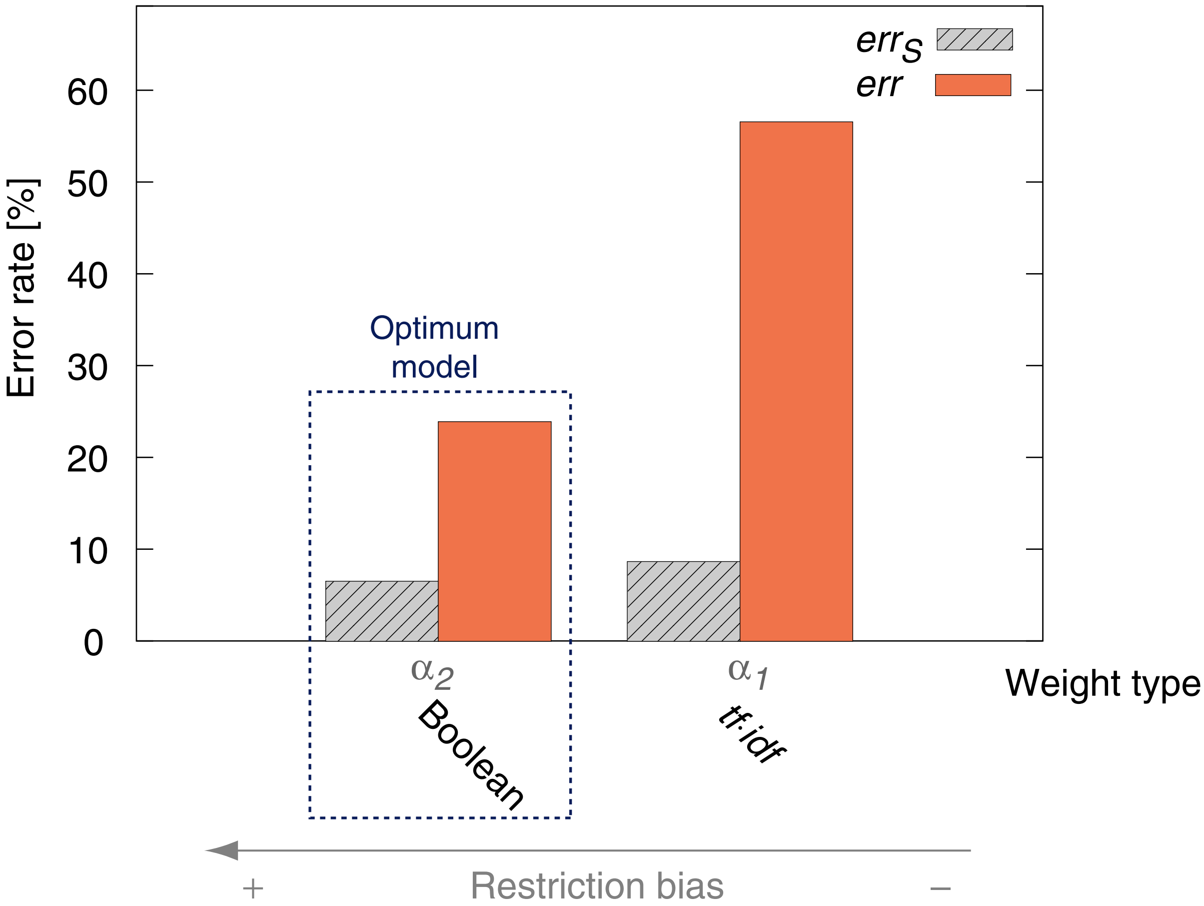
Robust Models in IR

Case Study I: Text Categorization



Robust Models in IR

Case Study I: Text Categorization

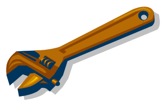


Robust Models in IR

Case Study II: Web Genre Classification

Given a web page, classify to one of the following 8 classes:

Shop

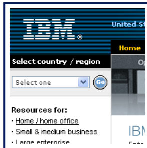


Help

Article

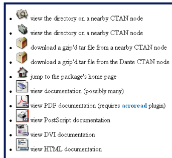
Discussion

Link collection



Non-pers. home

Personal home



Download

Experiment rationale:

- The sizes of existing genre corpora vary between 200 - 2500 documents.
- The number of the web genres in these corpora is between 3 and 16.
- The researchers report an very good (too good?) classification results.

→ The genre corpora are biased, e.g. because
  1. Editors collect their favored documents only.
  2. Editors introduce subconsciously correlations between topic and genre.

→ The classifiers that are learned with these corpora will not generalize well.

→ Learn two $h_{\alpha_1}, h_{\alpha_2}$ on corpus A and measure their *export accuracy* on corpus B.

Case Study II: Web Genre Classification

Experiment setup:

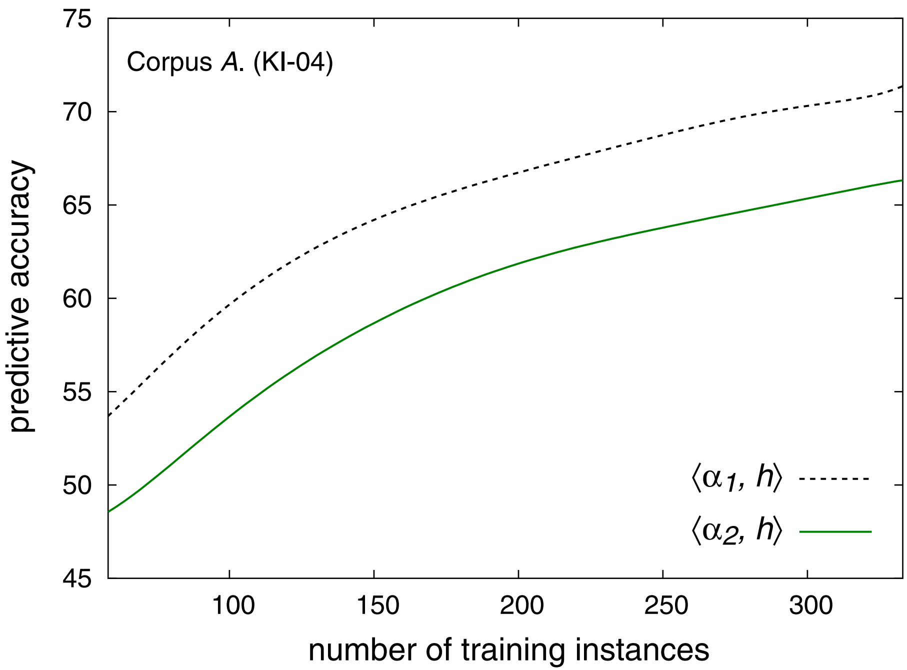
❑ Corpus A	KI-04, 1 200 documents
❑ Considered classes	article, discussion, shop, help, personal home, non-personal home, link collection, download
❑ Corpus B	7-Web-Genre, 900 documents
❑ Considered classes	listing (KI-04 link collection), eshop (KI-04 shop), home page (KI-04 personal home)
❑ Sample sizes	50-350, drawn i.i.d. from KI-04
❑ Inductive learner	SVM with linear kernel
❑ Model formation functions α	2 genre retrieval models

1. α_1: VSM-based model with 3 500 words
2. α_2: special concentrations measures plus core vocabulary (98 features)

Robust Models in IR

Case Study II: Web Genre Classification

Within corpus accuracy:

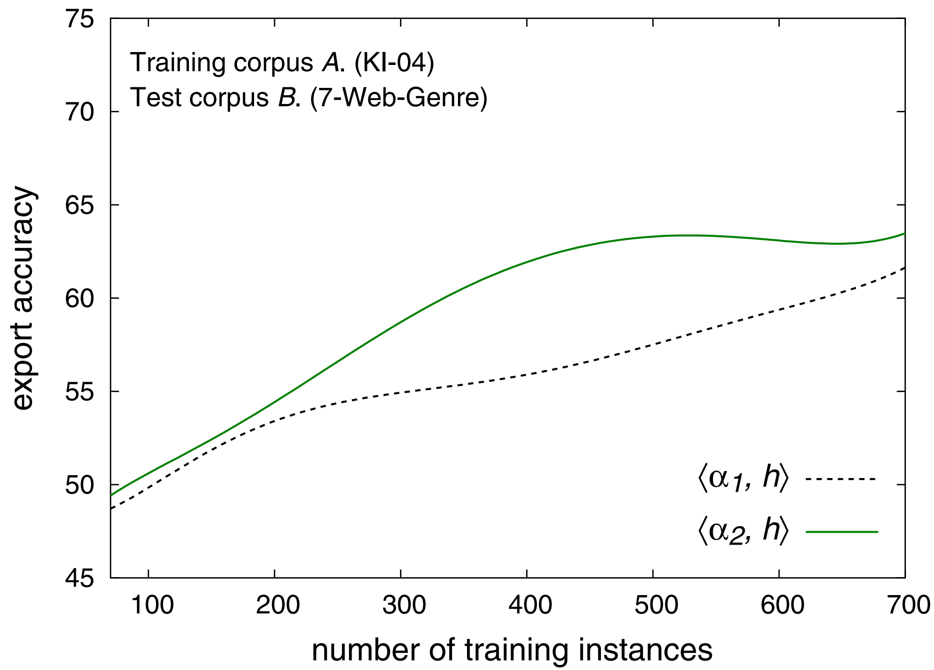


$$\mathit{err}_S(h^*_{\alpha_1}) < \mathit{err}_S(h^*_{\alpha_2})$$

Robust Models in IR

Case Study II: Web Genre Classification

Export accuracy:



$$\textit{err}(h^*_{\alpha_1}) > \textit{err}(h^*_{\alpha_2})$$

Summary

Summary

1. Be careful, if the ratio between sample size and application set (“test set”) becomes extreme:

   A model selection decision may not be justified by error minimization.

2. Consider …
   - a bias over-estimation of the less complex classifier or
   - a variance under-estimation of the more complex classifier.

3. In web scenarios the true error (generalization error) of a classifier cannot be analyzed:
   - ➜ develop a scale-up scenario to assess the impact on the error
   - ➜ if being in doubt stick to the less complex classifier

Thank you!

Excursus: Bias Types

Excursus: Bias Types

Bias in Classification Tasks