6. Supervised Techniques III (flipped)

DS-GA 1015, Text as Data
Arthur Spirling

March 16, 2021

Housekeeping

Housekeeping

HW1 being graded: looks good so far.

Housekeeping

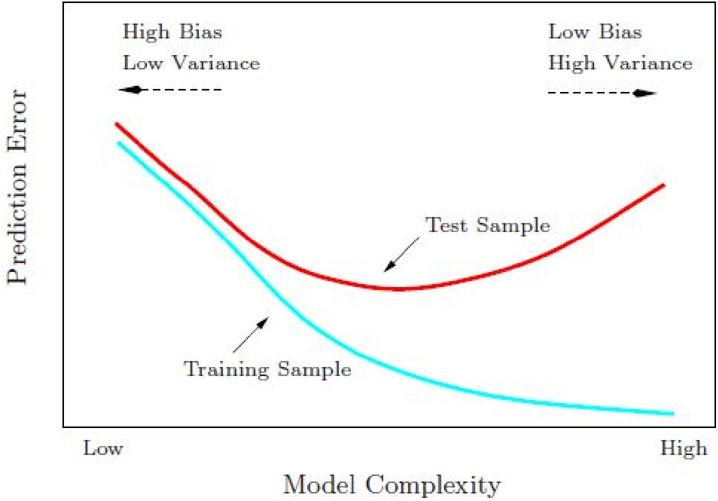
HW1 being graded: looks good so far.

HW2 out this week.

March 23: lab and flipped lecture will be in same session (no lab March 25).

Bias-Variance Tradeoff (Hastie et al, p38)

Bias-Variance Tradeoff (Hastie et al, p38)

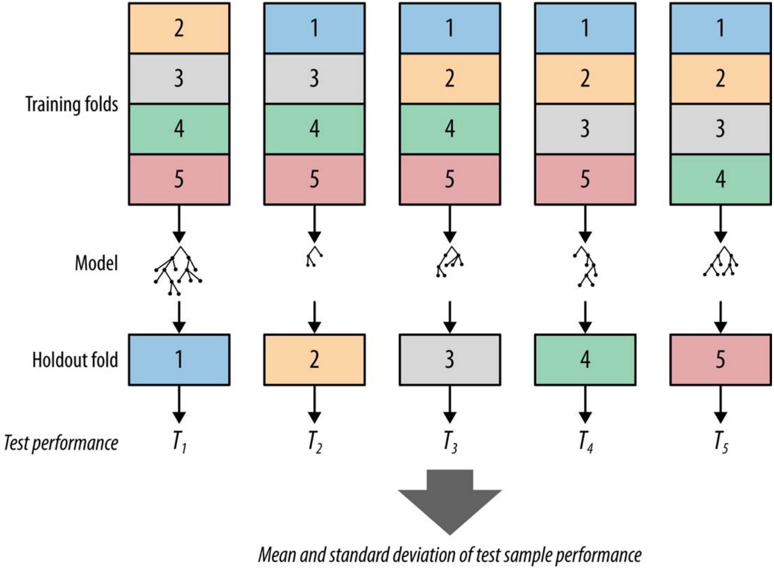


Moving left to right, what explains the training sample curve?

Moving left to right, what explains the test sample curve?

Graphically

Graphically



Rules of Thumb for Training/Test split

Rules of Thumb for Training/Test split

With little data (?), you are probably stuck with cross validation.

Rules of Thumb for Training/Test split

With little data (?), you are probably stuck with cross validation.

Otherwise, $\sim$ advice is random 70% (or 80%) for training,

Rules of Thumb for Training/Test split

With little data (?), you are probably stuck with cross validation.

Otherwise, $\sim$ advice is random 70% (or 80%) for training, 30% (or 20%) for testing.

Rules of Thumb for Training/Test split

With little data (?), you are probably stuck with cross validation.

Otherwise, $\sim$ advice is random 70% (or 80%) for training, 30% (or 20%) for testing.

Not unusual to also split training set into training and validation (80/20).

Rules of Thumb for Training/Test split

With little data (?), you are probably stuck with cross validation.

Otherwise, $\sim$ advice is random 70% (or 80%) for training, 30% (or 20%) for testing.

Not unusual to also split training set into training and validation (80/20). Idea is that validation set helps with tuning parameters to optimize model (e.g. number of variables, number of trees, number of splits in trees, different weighting schemes etc).

Rules of Thumb for Training/Test split

With little data (?), you are probably stuck with cross validation.

Otherwise, $\sim$ advice is random 70% (or 80%) for training, 30% (or 20%) for testing.

Not unusual to also split training set into training and validation (80/20). Idea is that validation set helps with tuning parameters to optimize model (e.g. number of variables, number of trees, number of splits in trees, different weighting schemes etc).

$\rightarrow$ test set ("hold out") is used to evaluate final performance of model.

Double Descent

Double Descent

We expect that higher complexity models have low bias but high variance. They overfit, and this gets worse as the model gets 'larger' (more parameters)

Double Descent

We expect that higher complexity models have low bias but high variance. They overfit, and this gets worse as the model gets 'larger' (more parameters)

But this does not appear to be true in (some) deep learning models:

Double Descent

We expect that higher complexity models have low bias but high variance. They overfit, and this gets worse as the model gets 'larger' (more parameters)

But this does not appear to be true in (some) deep learning models: as complexity increases, test error goes down, then up, then down again

Double Descent

We expect that higher complexity models have low bias but high variance. They overfit, and this gets worse as the model gets 'larger' (more parameters)

But this does not appear to be true in (some) deep learning models: as complexity increases, test error goes down, then up, then down again

$\rightarrow$ so-called double descent.

Double Descent

We expect that higher complexity models have low bias but high variance. They overfit, and this gets worse as the model gets 'larger' (more parameters)

But this does not appear to be true in (some) deep learning models: as complexity increases, test error goes down, then up, then down again

$\rightarrow$ so-called double descent. Occurs (briefly) after $p > n$, because there then you are picking the best fitting model from many (based on minimum norm),

Double Descent

We expect that higher complexity models have low bias but high variance. They overfit, and this gets worse as the model gets 'larger' (more parameters)

But this does not appear to be true in (some) deep learning models: as complexity increases, test error goes down, then up, then down again

$\rightarrow$ so-called double descent. Occurs (briefly) after $p > n$, because there then you are picking the best fitting model from many (based on minimum norm), so it is actually not as flexible/perfect a fit as at $p = n$.

Double Descent

We expect that higher complexity models have low bias but high variance. They overfit, and this gets worse as the model gets 'larger' (more parameters)

But this does not appear to be true in (some) deep learning models: as complexity increases, test error goes down, then up, then down again

$\rightarrow$ so-called double descent. Occurs (briefly) after $p > n$, because there then you are picking the best fitting model from many (based on minimum norm), so it is actually not as flexible/perfect a fit as at $p = n$.

$\Rightarrow$ lower variance

Double Descent (Nakkarin et al, 2019)

Exercise

Exercise

You are working for `yelp.com`.

Your boss wants to know what will predict reviews for restaurants in Greenwich Village in 2020, and gives you all the data from the Village in 2019.

Exercise

You are working for `yelp.com`.

Your boss wants to know what will predict reviews for restaurants in Greenwich Village in 2020, and gives you all the data from the Village in 2019.

She suggests you fit various (high dimensional) models to this data, and then pick the one with the lowest prediction error to fit to future data.

Exercise

You are working for `yelp.com`.

Your boss wants to know what will predict reviews for restaurants in Greenwich Village in 2020, and gives you all the data from the Village in 2019.

She suggests you fit various (high dimensional) models to this data, and then pick the one with the lowest prediction error to fit to future data. When you do so, you are unable to obtain anything like the same performance for the arriving 2020 data.

Exercise

You are working for `yelp.com`.

Your boss wants to know what will predict reviews for restaurants in Greenwich Village in 2020, and gives you all the data from the Village in 2019.

She suggests you fit various (high dimensional) models to this data, and then pick the one with the lowest prediction error to fit to future data. When you do so, you are unable to obtain anything like the same performance for the arriving 2020 data. Why?

Exercise

You are working for `yelp.com`.

Your boss wants to know what will predict reviews for restaurants in Greenwich Village in 2020, and gives you all the data from the Village in 2019.

She suggests you fit various (high dimensional) models to this data, and then pick the one with the lowest prediction error to fit to future data. When you do so, you are unable to obtain anything like the same performance for the arriving 2020 data. Why? What's wrong with this approach?

Exercise

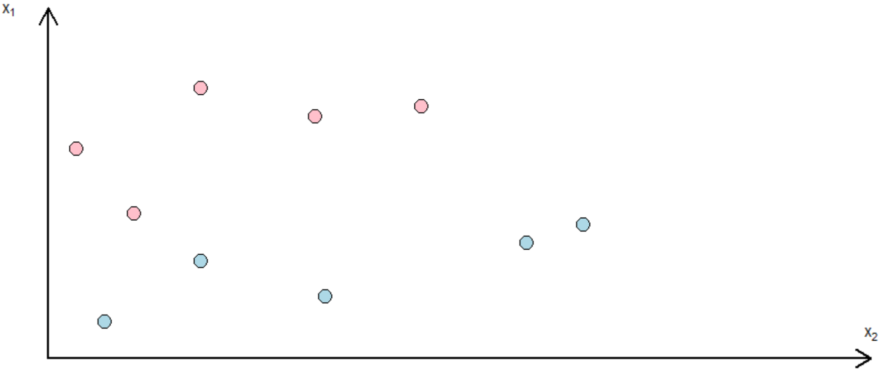
You are working for `yelp.com`.

Your boss wants to know what will predict reviews for restaurants in Greenwich Village in 2020, and gives you all the data from the Village in 2019.

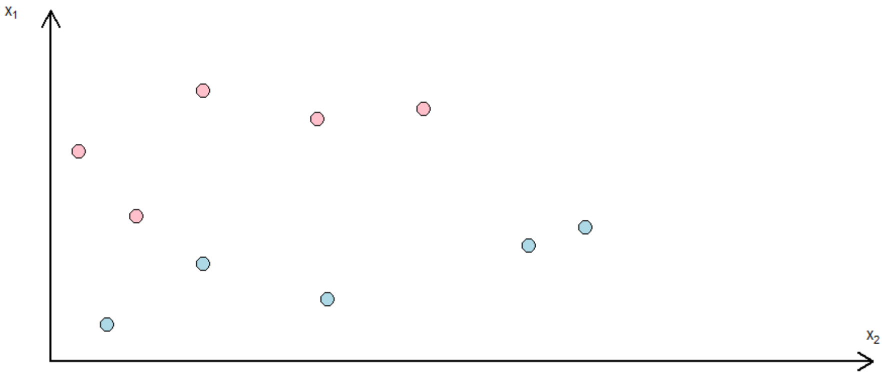
She suggests you fit various (high dimensional) models to this data, and then pick the one with the lowest prediction error to fit to future data. When you do so, you are unable to obtain anything like the same performance for the arriving 2020 data. Why? What's wrong with this approach?

The 10 Senators

The 10 Senators

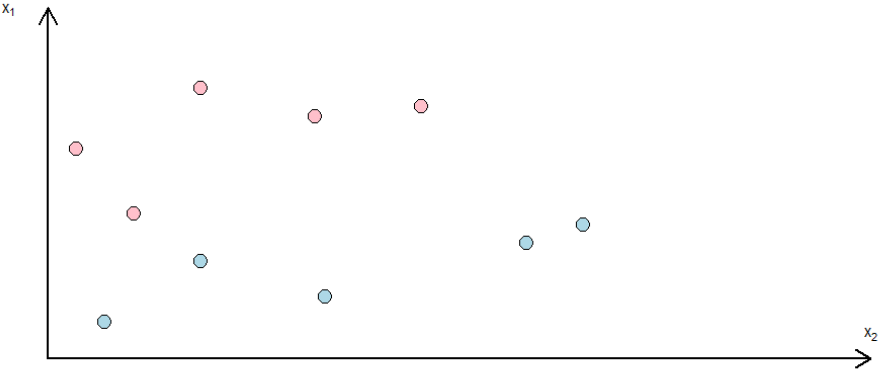


The 10 Senators



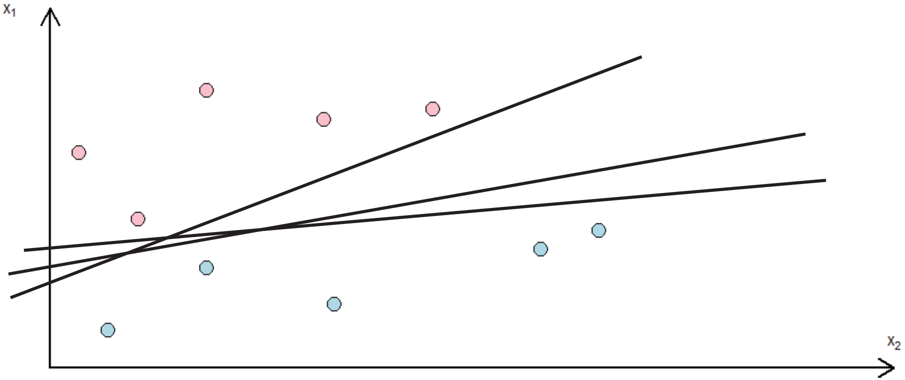
As the parties linearly separably?

The 10 Senators

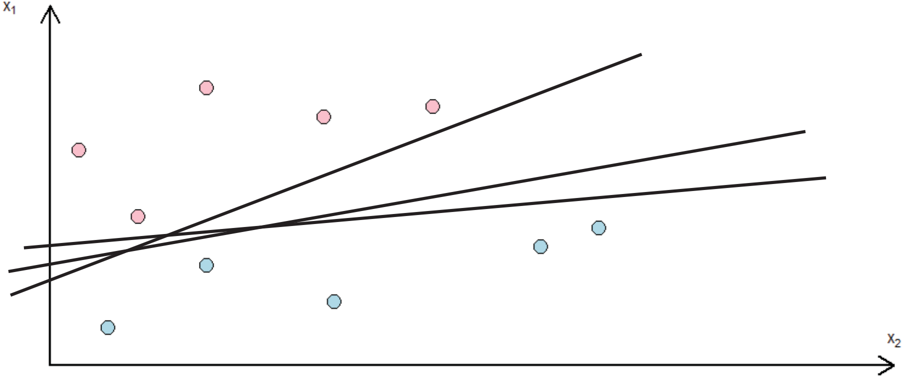


As the parties linearly separably? Where could you draw the line?

The 10 Senators

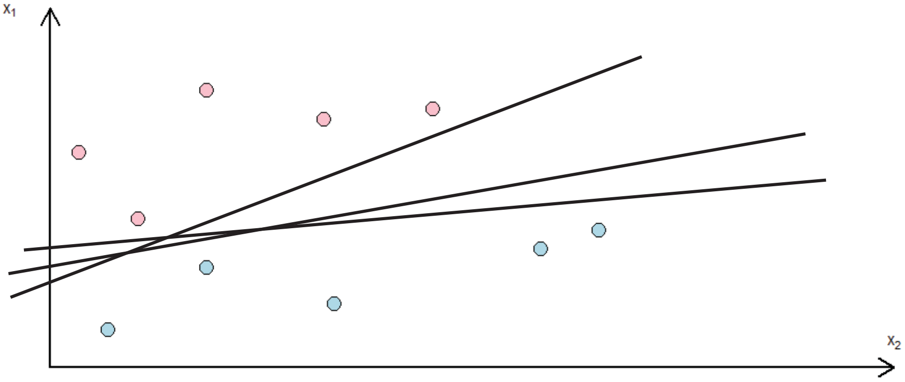


The 10 Senators



Which line should we prefer?

The 10 Senators



Which line should we prefer?

Exercise

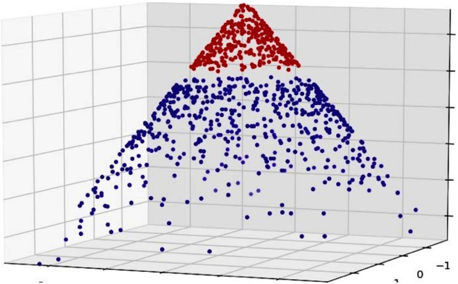
Exercise

Consider the figure.

It's a situation where each Senator's features are of three dimensions (rather than two).

How could we (optimally) separate the data in a linear way?

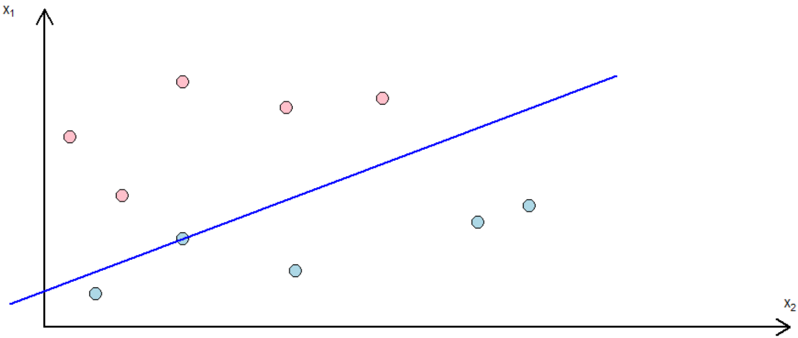
Can we still use a line?



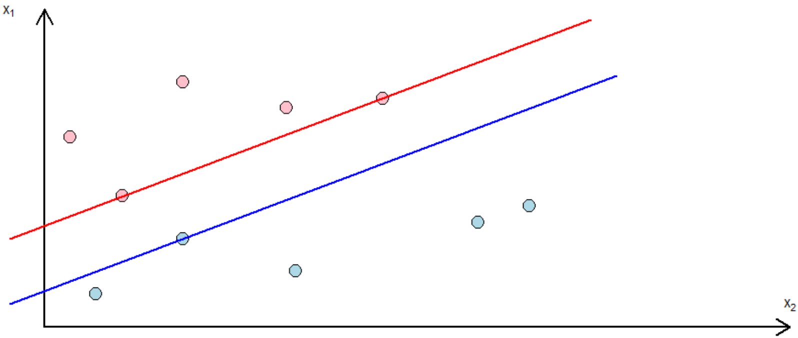
from http://www.edvancer.in/

Graphically...

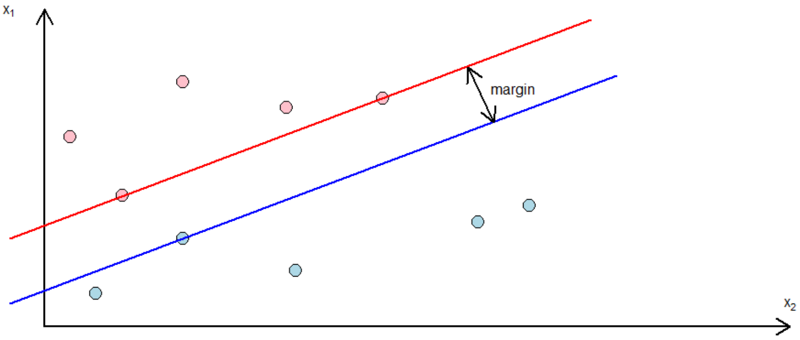
Graphically...



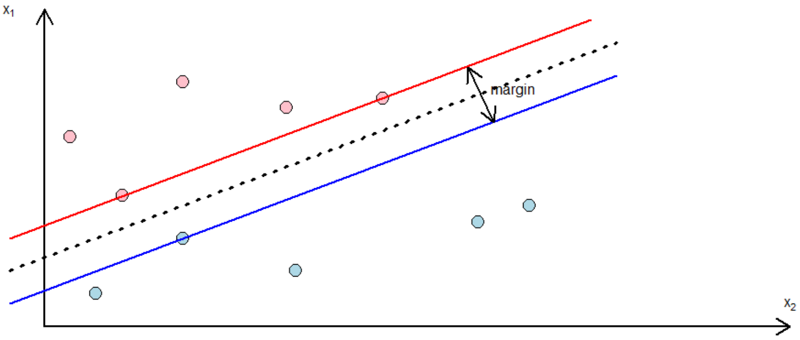
Graphically...



Graphically...



Graphically...



Questions

Questions

What is the motivation for wanting the maximum margin?

Questions

What is the motivation for wanting the maximum margin?

The optimal hyperplane must be half-way between the two parallel hyperplanes. Why?

Questions

What is the motivation for wanting the maximum margin?

The optimal hyperplane must be half-way between the two parallel hyperplanes. Why?

At least one Senator is one “their” line. Why?

Questions

What is the motivation for wanting the maximum margin?

The optimal hyperplane must be half-way between the two parallel hyperplanes. Why?

At least one Senator is one "their" line. Why?

What name do we give to the training examples on their respective hyperplanes?

Back to Diermeier et al, 2011

Back to Diermeier et al, 2011

Achieve 92% accuracy (!)

Back to Diermeier et al, 2011

Achieve 92% accuracy (!)

Sort words according to coefficients: very positive weights imply conservative words; very negative weights imply liberal words

Back to Diermeier et al, 2011

Achieve 92% accuracy (!)

Sort words according to coefficients: very positive weights imply conservative words; very negative weights imply liberal words. Argue that it is ‘values’ rather than economics that separates liberals from conservatives.

Back to Diermeier et al, 2011

Achieve 92% accuracy (!)

Sort words according to coefficients: very positive weights imply conservative words; very negative weights imply liberal words. Argue that it is 'values' rather than economics that separates liberals from conservatives.

Words			
Liberal		Conservative	
FAS: −199.49	SBA: −113.10	habeas: 193.55	homosexual: 103.07
Ethanol: −198.92	Nursing: −109.38	CFTC: 187.16	everglades: 102.87
Wealthiest: −159.74	Providence: −108.73	surtax: 151.81	tower: 101.67
Collider: −142.28	Arctic: −108.30	marriage: 145.79	tripartisan: 101.23
WIC: −140.14	Orange: −107.98	cloning: 141.71	PRC: 102.90
ILO: −139.89	Glaxo: −107.81	tritium: 133.49	scouts: 97.55
Handgun: −129.01	Libraries: −107.70	ranchers: 132.95	nashua: 99.32
Lobbyists: −128.95	Disabilities: −106.44	BTU: 121.92	ballistic: 97.22
Enron: −127.71	Prescription: −106.31	grazing: 121.59	salting: 94.28
Fishery: −127.30	NIH: −105.52	unfunded: 120.82	abortion: 91.94
Hydrogen: −122.59	Lobbying: −105.35	catfish: 120.82	NTSB: 93.81
Souter: −121.40	NRA: −105.20	IRS: 114.91	Haiti: 97.28
PTSD: −119.87	Trident: −104.15	unborn: 111.88	PAC: 92.85
Gun: −119.52	RNC: −103.46	Taiwan: 111.13	taxing: 90.39

Exercise

Exercise

Diermeier et al's results imply that liberals use '`Handgun`' more often than conservatives.

Exercise

Diermeier et al's results imply that liberals use '`Handgun`' more often than conservatives.

1 Does that imply that making conservative Senators use the word '`handgun`' more often will make them more liberal?

Exercise

Diermeier et al's results imply that liberals use '`Handgun`' more often than conservatives.

1. Does that imply that making conservative Senators use the word '`handgun`' more often will make them more liberal? What does your answer suggest about prediction vs explanation with supervised techniques?

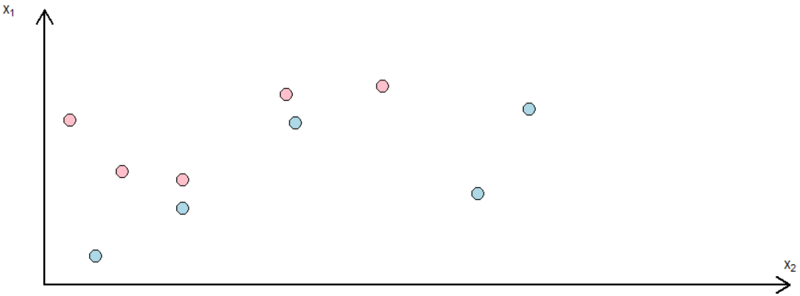
Exercise

Diermeier et al's results imply that liberals use '`Handgun`' more often than conservatives.

1 Does that imply that making conservative Senators use the word '`handgun`' more often will make them more liberal? What does your answer suggest about prediction vs explanation with supervised techniques?

2 what is the (most likely) problem in the causal claim that $X \rightarrow Y$ in the Diermeier et al study?

Oh dear...



What if...

What if. . .

The Senators were not linearly separable? ('soft margin' SVM problem)

What if...

The Senators were not linearly separable? ('soft margin' SVM problem)

Can introduce a hinge loss function into the minimization problem...

$$\boxed{\mathbb{L}(f(x), y) = \max(0, 1 - f(x)y)}$$

What if...

The Senators were not linearly separable? ('soft margin' SVM problem)

Can introduce a hinge loss function into the minimization problem...

$$\boxed{\mathbb{L}(f(x), y) = \max(0, 1 - f(x)y)}$$

$= 0$ if the xs are on the 'correct' side of the margin...

What if...

The Senators were not linearly separable? ('soft margin' SVM problem)

Can introduce a hinge loss function into the minimization problem...

$$\boxed{\mathbb{L}(f(x), y) = \max(0, 1 - f(x)y)}$$

$= 0$ if the xs are on the 'correct' side of the margin...

And proportional to the distance from the margin *if* the point is on the 'wrong' side of the margin.

What if. . .

The Senators were not linearly separable? ('soft margin' SVM problem)

Can introduce a hinge loss function into the minimization problem. . .

$$\boxed{\mathbb{L}(f(x), y) = \max(0, 1 - f(x)y)}$$

$= 0$ if the xs are on the 'correct' side of the margin. . .

And proportional to the distance from the margin *if* the point is on the 'wrong' side of the margin.

Hyperplane(s) will be drawn in way that is more sensitive to 'bigger' mistakes in classification.

Exercise

Exercise

We want $\mathbf{w} \cdot \mathbf{x} - b \geq 1$, if $y_i =$ Republican

Exercise

We want $\mathbf{w} \cdot \mathbf{x} - b \geq 1$, if $y_i =$ Republican

and $\mathbf{w} \cdot \mathbf{x} - b \leq -1$, if $y_i =$ Democrat

Exercise

We want $\mathbf{w} \cdot \mathbf{x} - b \geq 1$, if $y_i =$ Republican

and $\mathbf{w} \cdot \mathbf{x} - b \leq -1$, if $y_i =$ Democrat

Another way (slightly different notation) to express this is that we want

Exercise

We want $\mathbf{w} \cdot \mathbf{x} - b \geq 1$, if $y_i =$ Republican

and $\mathbf{w} \cdot \mathbf{x} - b \leq -1$, if $y_i =$ Democrat

Another way (slightly different notation) to express this is that we want

$$f(x_i) = \begin{cases} \geq +1 \text{ if } y_i = +1 \\ \leq -1 \text{ if } y_i = -1 \end{cases}$$

Exercise

We want $\mathbf{w} \cdot \mathbf{x} - b \geq 1$, if $y_i =$ Republican
and $\mathbf{w} \cdot \mathbf{x} - b \leq -1$, if $y_i =$ Democrat

Another way (slightly different notation) to express this is that we want

$$f(x_i) = \begin{cases} \geq +1 \text{ if } y_i = +1 \\ \leq -1 \text{ if } y_i = -1 \end{cases}$$

1 If a classification is correct,

Exercise

We want $\mathbf{w} \cdot \mathbf{x} - b \geq 1$, if $y_i =$ Republican

and $\mathbf{w} \cdot \mathbf{x} - b \leq -1$, if $y_i =$ Democrat

Another way (slightly different notation) to express this is that we want

$$f(x_i) = \begin{cases} \geq +1 \text{ if } y_i = +1 \\ \leq -1 \text{ if } y_i = -1 \end{cases}$$

1 If a classification is correct, what do we know about the value of $y_i \times f(x_i)$?

Exercise

We want $\mathbf{w} \cdot \mathbf{x} - b \geq 1$, if $y_i =$ Republican
and $\mathbf{w} \cdot \mathbf{x} - b \leq -1$, if $y_i =$ Democrat

Another way (slightly different notation) to express this is that we want

$$f(x_i) = \begin{cases} \geq +1 \text{ if } y_i = +1 \\ \leq -1 \text{ if } y_i = -1 \end{cases}$$

1. If a classification is correct, what do we know about the value of $y_i \times f(x_i)$?
2. If a Republican is misclassified as a Democrat, what value will $\mathbb{L}(f(x), y) = \max(0, 1 - f(x)y)$ take?

Exercise

We want $\mathbf{w} \cdot \mathbf{x} - b \geq 1$, if $y_i =$ Republican
and $\mathbf{w} \cdot \mathbf{x} - b \leq -1$, if $y_i =$ Democrat

Another way (slightly different notation) to express this is that we want

$$f(x_i) = \begin{cases} \geq +1 \text{ if } y_i = +1 \\ \leq -1 \text{ if } y_i = -1 \end{cases}$$

1. If a classification is correct, what do we know about the value of $y_i \times f(x_i)$?
2. If a Republican is misclassified as a Democrat, what value will $\mathbb{L}(f(x), y) = \max(0, 1 - f(x)y)$ take?
3. Two Republicans were misclassified as Democrats by the machine. In the first case, $f(x_i) = -2$. In the second case, $f(x_i) = -100$. Which has the 'worse' value of hinge loss?

Solutions

Solutions

Another way to express this is that we want

Solutions

Another way to express this is that we want

$$f(x_i) = \begin{cases} \geq +1 \text{ if } y_i = +1 \\ \leq -1 \text{ if } y_i = -1 \end{cases}$$

Solutions

Another way to express this is that we want

$$f(x_i) = \begin{cases} \geq +1 \text{ if } y_i = +1 \\ \leq -1 \text{ if } y_i = -1 \end{cases}$$

1 If a classification is correct,

Solutions

Another way to express this is that we want

$$f(x_i) = \begin{cases} \geq +1 \text{ if } y_i = +1 \\ \leq -1 \text{ if } y_i = -1 \end{cases}$$

1 If a classification is correct, what do we know about the value of $y_i \times f(x_i)$?

Solutions

Another way to express this is that we want

$$f(x_i) = \begin{cases} \geq +1 \text{ if } y_i = +1 \\ \leq -1 \text{ if } y_i = -1 \end{cases}$$

1 If a classification is correct, what do we know about the value of $y_i \times f(x_i)$?

Positive, ≥ 1.

Solutions

Another way to express this is that we want

$$f(x_i) = \begin{cases} \geq +1 \text{ if } y_i = +1 \\ \leq -1 \text{ if } y_i = -1 \end{cases}$$

1. If a classification is correct, what do we know about the value of $y_i \times f(x_i)$?

   Positive, ≥ 1.

2. If a Republican is misclassified as a Democrat, what value will $\mathbb{L}(f(x), y) = \max(0, 1 - f(x)y)$ take?

Solutions

Another way to express this is that we want

$$f(x_i) = \begin{cases} \geq +1 \text{ if } y_i = +1 \\ \leq -1 \text{ if } y_i = -1 \end{cases}$$

1 If a classification is correct, what do we know about the value of $y_i \times f(x_i)$?

Positive, ≥ 1.

2 If a Republican is misclassified as a Democrat, what value will $\mathbb{L}(f(x), y) = \max(0, 1 - f(x)y)$ take?

Now, $f(x)$ is negative ('looks like' a Dem), but y is positive,

Solutions

Another way to express this is that we want

$$f(x_i) = \begin{cases} \geq +1 \text{ if } y_i = +1 \\ \leq -1 \text{ if } y_i = -1 \end{cases}$$

1 If a classification is correct, what do we know about the value of $y_i \times f(x_i)$?

Positive, ≥ 1.

2 If a Republican is misclassified as a Democrat, what value will $\mathbb{L}(f(x), y) = \max(0, 1 - f(x)y)$ take?

Now, $f(x)$ is negative ('looks like' a Dem), but y is positive, so $f(x)y$ is negative.

Solutions

Another way to express this is that we want

$$f(x_i) = \begin{cases} \geq +1 \text{ if } y_i = +1 \\ \leq -1 \text{ if } y_i = -1 \end{cases}$$

1. If a classification is correct, what do we know about the value of $y_i \times f(x_i)$?

   Positive, ≥ 1.

2. If a Republican is misclassified as a Democrat, what value will $\mathbb{L}(f(x), y) = \max(0, 1 - f(x)y)$ take?

   Now, $f(x)$ is negative ('looks like' a Dem), but y is positive, so $f(x)y$ is negative. But then $1 - f(x)y$ is large, and is the maximum of the set.

Solutions

Another way to express this is that we want

$$f(x_i) = \begin{cases} \geq +1 \text{ if } y_i = +1 \\ \leq -1 \text{ if } y_i = -1 \end{cases}$$

1. If a classification is correct, what do we know about the value of $y_i \times f(x_i)$?

   Positive, ≥ 1.

2. If a Republican is misclassified as a Democrat, what value will $\mathbb{L}(f(x), y) = \max(0, 1 - f(x)y)$ take?

   Now, $f(x)$ is negative ('looks like' a Dem), but y is positive, so $f(x)y$ is negative. But then $1 - f(x)y$ is large, and is the maximum of the set.

3. Two Republicans were misclassified as Democrats by the machine. In the first case, $f(x_i) = -2$. In the second case, $f(x_i) = -100$. Which has the 'worse' value of hinge loss?

Solutions

Another way to express this is that we want

$$f(x_i) = \begin{cases} \geq +1 \text{ if } y_i = +1 \\ \leq -1 \text{ if } y_i = -1 \end{cases}$$

1. If a classification is correct, what do we know about the value of $y_i \times f(x_i)$?

   Positive, ≥ 1.

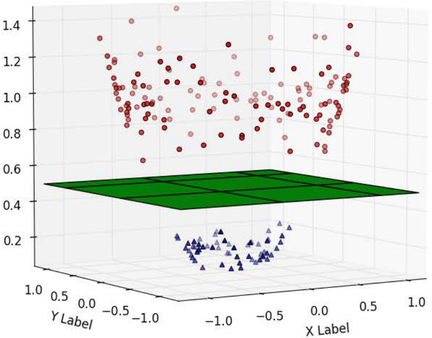
2. If a Republican is misclassified as a Democrat, what value will $\mathbb{L}(f(x), y) = \max(0, 1 - f(x)y)$ take?

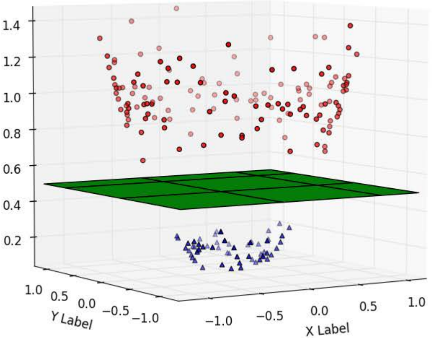
   Now, $f(x)$ is negative ('looks like' a Dem), but y is positive, so $f(x)y$ is negative. But then $1 - f(x)y$ is large, and is the maximum of the set.

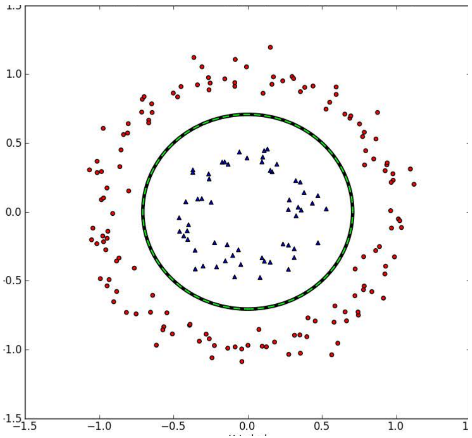
3. Two Republicans were misclassified as Democrats by the machine. In the first case, $f(x_i) = -2$. In the second case, $f(x_i) = -100$. Which has the 'worse' value of hinge loss?

   $1 - f(x)y$ is $1 - (-2)(+1)$ in first case and $1 - (-100)(+1)$ in second case. Hinge loss larger in second case!

Kernels


1.4
1.2
1.0
0.8
0.6
0.4
0.2
1.0 0.5 0.0 −0.5 −1.0
Y Label
−1.0 −0.5 0.0 0.5 1.0
X Label





from www.eric-kim.net

Kernel Methods

Kernel Methods

Explicitly transforming data into a new space can be very expensive in terms of computation.

Kernel Methods

Explicitly transforming data into a new space can be very expensive in terms of computation.

But What if we could do the transformation, and take the distances between observations implicitly, and use them for our classification?

Kernel Methods

Explicitly transforming data into a new space can be very expensive in terms of computation.

But What if we could do the transformation, and take the distances between observations implicitly, and use them for our classification?

$\rightarrow$ exactly what kernel methods do:

Kernel Methods

Explicitly transforming data into a new space can be very expensive in terms of computation.

But What if we could do the transformation, and take the distances between observations implicitly, and use them for our classification?

$\rightarrow$ exactly what kernel methods do: use kernel functions,

Kernel Methods

Explicitly transforming data into a new space can be very expensive in terms of computation.

But What if we could do the transformation, and take the distances between observations implicitly, and use them for our classification?

$\rightarrow$ exactly what kernel methods do: use kernel functions, $K(\mathbf{x}_i, \mathbf{x}_j)$ to compute the i, j pairwise dot products for training data *as if* it had been transformed.

Kernel Methods

Explicitly transforming data into a new space can be very expensive in terms of computation.

But What if we could do the transformation, and take the distances between observations implicitly, and use them for our classification?

→ exactly what kernel methods do: use kernel functions, $K(\mathbf{x}_i, \mathbf{x}_j)$ to compute the i, j pairwise dot products for training data *as if* it had been transformed. Can then feed those inner products to the classifier.

Kernel Methods

Explicitly transforming data into a new space can be very expensive in terms of computation.

But What if we could do the transformation, and take the distances between observations implicitly, and use them for our classification?

$\rightarrow$ exactly what kernel methods do: use kernel functions, $K(\mathbf{x}_i, \mathbf{x}_j)$ to compute the i, j pairwise dot products for training data *as if* it had been transformed. Can then feed those inner products to the classifier.

this 'kernel trick' cuts cost considerably,

Kernel Methods

Explicitly transforming data into a new space can be very expensive in terms of computation.

But What if we could do the transformation, and take the distances between observations implicitly, and use them for our classification?

$\rightarrow$ exactly what kernel methods do: use kernel functions, $K(\mathbf{x}_i, \mathbf{x}_j)$ to compute the i, j pairwise dot products for training data *as if* it had been transformed. Can then feed those inner products to the classifier.

this 'kernel trick' cuts cost considerably, though choosing and tuning the 'correct' kernel may be difficult.

Kernel Methods

Explicitly transforming data into a new space can be very expensive in terms of computation.

But What if we could do the transformation, and take the distances between observations implicitly, and use them for our classification?

$\rightarrow$ exactly what kernel methods do: use kernel functions, $K(\mathbf{x}_i, \mathbf{x}_j)$ to compute the i, j pairwise dot products for training data *as if* it had been transformed. Can then feed those inner products to the classifier.

this 'kernel trick' cuts cost considerably, though choosing and tuning the 'correct' kernel may be difficult.

For text analysis,

Kernel Methods

Explicitly transforming data into a new space can be very expensive in terms of computation.

But What if we could do the transformation, and take the distances between observations implicitly, and use them for our classification?

$\rightarrow$ exactly what kernel methods do: use kernel functions, $K(\mathbf{x}_i, \mathbf{x}_j)$ to compute the i, j pairwise dot products for training data *as if* it had been transformed. Can then feed those inner products to the classifier.

this 'kernel trick' cuts cost considerably, though choosing and tuning the 'correct' kernel may be difficult.

For text analysis, string kernels use a function $K(a, b)$ to implicity calculate the distance between strings of characters via the number of subsequences they have in common.

Kernel Methods in Action
(Self-indulgent Slides on Spirling, 2011)

Overview

Overview

Overview

"establishing a firm and permanent friendship. . ."

Overview

“establishing a firm and permanent friendship. . . ”

“United States acknowledge the lands reserved to the Oneida. . . ”

Overview

“establishing a firm and permanent friendship. . .”

“United States acknowledge the lands reserved to the Oneida. . .”

Overview

“establishing a firm and permanent friendship. . . ”

“United States acknowledge the lands reserved to the Oneida. . . ”

“That the President is hereby authorized and required, whenever in his opinion any reservation of such Indians. . . to allot the lands in said reservation. . . ”

Question

- terms of treaties got worse over time. . .

Question

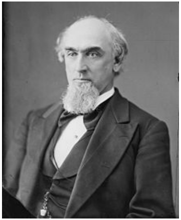
- terms of treaties got worse over time...
- but 1871 saw change in treaty making initiative:
  - previously purview of President under Article II

Question

- terms of treaties got worse over time. . .
- but 1871 saw change in treaty making initiative:
  - previously purview of President under Article II
  - treaties became Congressional 'bill-like' entities

Question

- terms of treaties got worse over time. . .
- but 1871 saw change in treaty making initiative:
  - previously purview of President under Article II
  - treaties became Congressional ‘bill-like’ entities
  - presumably different incentives (?), but do treaties look much different?



What we have

What we have

- ‘Valid and operable’ (365):
  1784–1868, ratified under Article
  II.

What we have

- ‘Valid and operable’ (365): 1784–1868, ratified under Article II.
- ‘Ratified Agreements’ (77): post-1871 after purported ‘end’ of treaty-making, ratified in *statute* form.

What we have

- 'Valid and operable' (365): 1784–1868, ratified under Article II.
- 'Ratified Agreements' (77): post-1871 after purported 'end' of treaty-making, ratified in *statute* form.
- 'Rejected by Congress' (85): submitted to the Senate, but not ratified.

What we have

- 'Valid and operable' (365): 1784–1868, ratified under Article II.
- 'Ratified Agreements' (77): post-1871 after purported 'end' of treaty-making, ratified in *statute* form.
- 'Rejected by Congress' (85): submitted to the Senate, but not ratified.
- 'Unratified Treaties' (68): signed before 1868 yet never submitted for Senate ratification.

What we have

- 'Valid and operable' (365): 1784–1868, ratified under Article II.
- 'Ratified Agreements' (77): post-1871 after purported 'end' of treaty-making, ratified in *statute* form.
- 'Rejected by Congress' (85): submitted to the Senate, but not ratified.
- 'Unratified Treaties' (68): signed before 1868 yet never submitted for Senate ratification.

A Treaty of Limits between the United States of America and the Chaktaw [sic] nation of Indians.

THOMAS JEFFERSON, President of the United States of America, by James Robertson, of Tennessee, and Silas Dinsmoor, of New Hampshire, agent of the United States to the Chaktaws, commissioners plenipotentiary of the United States, on the one part, and the Mingoes, Chiefs and warriors of the Chaktaw nation of Indians, in council assembled, on the other part, have entered into the following agreement, viz:

ARTICLE 1.

The Mingoes, chiefs and warriors of the Choctaw nation of Indians in behalf of themselves, and the said nation, do by these presents cede to the United States of America, all the lands to which they now have or ever had claim, lying to the right of the following lines, to say. Beginning at a branch of the Humacheeto where the same is intersected by the present Choctaw boundary, and also by the path leading from Natchez to the county of Washington, usually called M'Clarey's path, thence eastwardly along M'Clarey's path, to the east or left bank of Pearl river thence on such a direct line as would touch the lower end of a bluff on the left bank of Chickasawhay river the first above the Hiyoowannee towns, called Broken Bluff, to a point within four miles of the Broken Bluff, thence in a direct line nearly parallel with the river to a point whence an east line of four miles in length will intersect the river below the lowest

ARTICLE 4.

The Mingoes, chiefs, and warriors of the Choctaws, certify that a tract of land not exceeding fifteen hundred acres, situated between the Tombigbee river and Jackson's creek, the front or river line extending down the river from a blazed white oak standing on the left bank of the Tombigbee near the head of the shoal, next above Hobukentoopa, and claimed by John M'Grew was in fact granted to the said M'Grew by Opiomingo Hesnitta, and others, many years ago, and they respectfully request the government of the United States to establish the claim of the said M'Grew to the said fifteen hundred acres.

Done on Mount Dexter, in Pooshapukanuk, in the Choctaw country, this sixteenth day of November, in the year of our Lord one thousand eight hundred and five, and of the independence of the United States of America the thirtieth.

Commissioners:

James Robertson, [L. S.

Silas Dinsmoor, [L. S.]

Great Medal Mingos:

Pukshunnubbee. his x mark, [L. S.]

Mingo Hoomastubbee, his x mark, [L. S.]

Pooshamattaha, his x mark, [L. S.]

Chiefs and warriors:

Ookchummee, his x mark, [L. S.]

Tuskamiubbee, his x mark, [L. S.]

What we want

What we want

1. capture general patterns, via data reduction

What we want

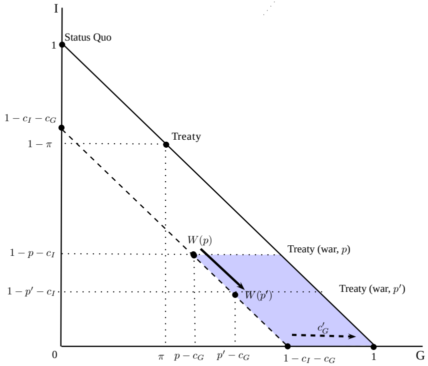
1. capture general patterns, via data reduction
2. report number of 'dimensions'

What we want

1. capture general patterns, via data reduction
2. report number of ‘dimensions’
3. report nature of ‘dimensions’

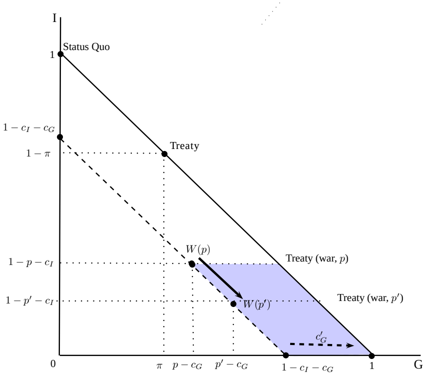
What we want

1. capture general patterns, via data reduction
2. report number of 'dimensions'
3. report nature of 'dimensions'
4. place documents on contract curve (if appropriate)



What we want

1. capture general patterns, via data reduction
2. report number of 'dimensions'
3. report nature of 'dimensions'
4. place documents on contract curve (if appropriate)



What we do

- put data in numerical form

What we do

- put data in numerical form
- scale using (some kind of) components/coordinate analysis

What we do

- string kernels
  - in sense of Lodhi et al

- put data in numerical form
- scale using (some kind of) components/coordinate analysis

What we do

- string kernels
  - in sense of Lodhi et al
  - break words into substrings of some length k (4–7)

- put data in numerical form
- scale using (some kind of) components/coordinate analysis

What we do

- put data in numerical form
- scale using (some kind of) components/coordinate analysis

- string kernels
  - in sense of Lodhi et al
  - break words into substrings of some length k (4–7)
  - calculate inner-product of two documents as (normalized) number of substrings in common

What we do

- put data in numerical form
- scale using (some kind of) components/coordinate analysis

- string kernels
  - in sense of Lodhi et al
  - break words into substrings of some length k (4–7)
  - calculate inner-product of two documents as (normalized) number of substrings in common
  - work on $d \times d$ kernel matrix

What we do

- put data in numerical form
- scale using (some kind of) components/coordinate analysis

- string kernels
  - in sense of Lodhi et al
  - break words into substrings of some length k (4–7)
  - calculate inner-product of two documents as (normalized) number of substrings in common
  - work on $d \times d$ kernel matrix
  - fairly fast, preserves text structure, outperforms TDMs in classification tests

What we do

- put data in numerical form
- scale using (some kind of) components/coordinate analysis

- string kernels
  - in sense of Lodhi et al
  - break words into substrings of some length k (4–7)
  - calculate inner-product of two documents as (normalized) number of substrings in common
  - work on $d \times d$ kernel matrix
  - fairly fast, preserves text structure, outperforms TDMs in classification tests
- kernel PCA

What we do

- put data in numerical form
- scale using (some kind of) compo-nents/coordinate analysis

- string kernels
  - in sense of Lodhi et al
  - break words into substrings of some length k (4–7)
  - calculate inner-product of two documents as (normalized) number of substrings in common
  - work on $d \times d$ kernel matrix
  - fairly fast, preserves text structure, outperforms TDMs in classification tests
- kernel PCA
  - in sense of Schoelkopf et al

What we do

- put data in numerical form
- scale using (some kind of) components/coordinate analysis

- string kernels
  - in sense of Lodhi et al
  - break words into substrings of some length k (4–7)
  - calculate inner-product of two documents as (normalized) number of substrings in common
  - work on $d \times d$ kernel matrix
  - fairly fast, preserves text structure, outperforms TDMs in classification tests
- kernel PCA
  - in sense of Schoelkopf et al
  - as PCA, but components may be non-linear combinations of original variables

What we do

- put data in numerical form
- scale using (some kind of) components/coordinate analysis

- string kernels
  - in sense of Lodhi et al
  - break words into substrings of some length k (4–7)
  - calculate inner-product of two documents as (normalized) number of substrings in common
  - work on $d \times d$ kernel matrix
  - fairly fast, preserves text structure, outperforms TDMs in classification tests
- kernel PCA
  - in sense of Schoelkopf et al
  - as PCA, but components may be non-linear combinations of original variables
  - (relative) GoF available via eigenvalues

1. `peace not war between`
2. `brothers not warfare now`
3. `be war not friendship`

documents are similar in word use terms. . .

1. `peace not war between`
2. `brothers not warfare now`
3. `be war not friendship`

documents are similar in word use terms. . .

but (1) and (2) share more substrings (of length 4):

1. `peace not war between`
2. `brothers not warfare now`
3. `be war not friendship`

not w,

1. peace not war between
2. brothers not warfare now
3. be war not friendship

not w,

1. `peace not war between`
2. `brothers not warfare now`
3. `be war not friendship`

not w,

1. peace |not w|ar between
2. brothers |not w|arfare now
3. be war not friendship

not w,

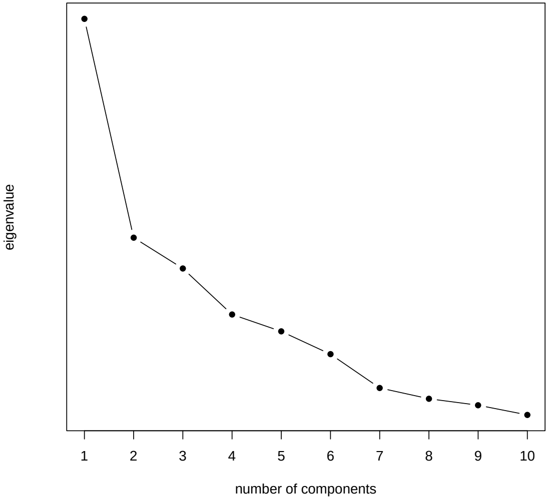
1. peace n|ot wa|r between
2. brothers n|ot wa|rfare now
3. be war not friendship

ot wa,

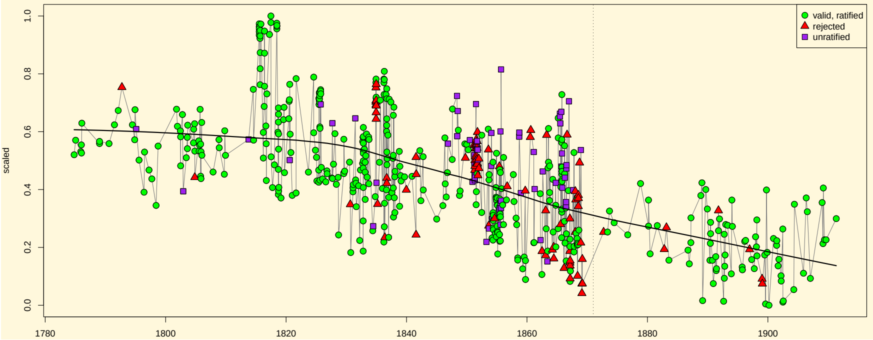
1. peace no|t war|between
2. brothers no|t war|fare now
3. be war not friendship

t war

What we get, I



What we get, II



Exercise

Exercise

Using the ideas we discussed at the start of lecture, how should one go about picking a kernel (from the large variety on offer) for the problem at hand?