Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**

Reduction to Classification

Jordan Boyd-Graber
University of Colorado Boulder
LECTURE 13

Busy Week

Content Questions

Content Questions

Content Questions

Content Questions

Content Questions

Content Questions

Content Questions

Content Questions

Content Questions

Content Questions

.

Administrivia

- How is the course going?
- What do you like?
- What don't you like?
- What should we do for an undergrad section?

Administrivia

- Boosting Due on Friday
- Midterm Next Week: 1.5 Hours
- Project Meetings
- Default Project

Default Project

```
id,question,correctAnswer,answerA,answerB,answerC,answerD
196635,This phenomenon occurs twice as quickly in the Moran model as in the Wright-Fisher model.,A,Genetic drift,Hamiltonian (quar
207540,"Coenocytic members of this kingdom lack septa, and another group in this kingdom, the ascomycetes, reproduces asexually us
99,This painting was drawn from a cartoon by Michelangelo showing two of the title characters kissing.,B,Hamiltonian (quantum mech
196778,"This geological process saw increased activity during the mid-Cretaceous in its namesake ""pulse."" The Vine-Matthews-Morl
207564,This entity is equal to the curl of the vector potential.,D,Moment of inertia,Angular momentum,Electric field,Magnetic fiel
204696,A mathematical ring consists of a set and this many operations.,B,Georg Wilhelm Friedrich Hegel,2 (number),Hamiltonian (qua
196848,"The Hoecht stain binds to the minor groove of this molecule, forming G-quadruplexes.",A,DNA,Adenosine triphosphate,Cyclic
196854,"This substance's ability to form metal ligands is exemplified in the coordination complex it forms with copper and water,
```

Defining a Code Book

- Want to decide whether a name is German, Argentine, or Chinese
- Using ECOC
- What do we need first?

Defining a Code Book

- Want to decide whether a name is German, Argentine, or Chinese
- Using ECOC
- What do we need first?

Class	b_1	b_2	b_3	b_4
Chinese	1	0	0	1
German	0	0	1	0
Argentine	1	1	1	0

Training Data

German	Chinese	Argentine
Mann	Cao Xueqin	Puig
Goethe	Lu Xun	Borges
Grass	Gao Xingjian	Cortazar

Training Data

German	Chinese	Argentine
Mann	Cao Xueqin	Puig
Goethe	Lu Xun	Borges
Grass	Gao Xingjian	Cortazar

What are the training examples for each classifier?

Training Data

German	Chinese	Argentine
Mann	Cao Xueqin	Puig
Goethe	Lu Xun	Borges
Grass	Gao Xingjian	Cortazar

Class	b_1	b_2	b_3	b_4
Chinese	1	0	0	1
German	0	0	1	0
Argentine	1	1	1	0

What are the training examples for each classifier?

Training Examples

	h_1	h_2	h_3	h_4

Training Examples

	h_1	h_2	h_3	h_4
Mann	-	-	+	-
Goethe	-	-	+	-
Grass	-	-	+	-

Training Examples

	h_1	h_2	h_3	h_4
Mann	-	-	+	-
Goethe	-	-	+	-
Grass	-	-	+	-
Cao Xue	+	-	-	+
Lu Xun	+	-	-	+
Gao Xingjian	+	-	-	+

Training Examples

	h_1	h_2	h_3	h_4
Mann	-	-	+	-
Goethe	-	-	+	-
Grass	-	-	+	-
Cao Xue	+	-	-	+
Lu Xun	+	-	-	+
Gao Xingjian	+	-	-	+
Puig	+	+	+	-
Borges	+	+	+	-
Cortazar	+	+	+	-

Classification

Class	b_1	b_2	b_3	b_4
Chinese	1	0	0	1
German	0	0	1	0
Argentine	1	1	1	0

Classification

Class	b_1	b_2	b_3	b_4
Chinese	1	0	0	1
German	0	0	1	0
Argentine	1	1	1	0

- $(0, 0, 0, 1) \rightarrow$
-

Classification

Class	b_1	b_2	b_3	b_4
Chinese	1	0	0	1
German	0	0	1	0
Argentine	1	1	1	0

- $(0, 0, 0, 1) \rightarrow$ German
-

Classification

Class	b_1	b_2	b_3	b_4
Chinese	1	0	0	1
German	0	0	1	0
Argentine	1	1	1	0

- $(0, 0, 0, 1) \rightarrow$ German
- $(0, 1, 0, 1) \rightarrow$

Classification

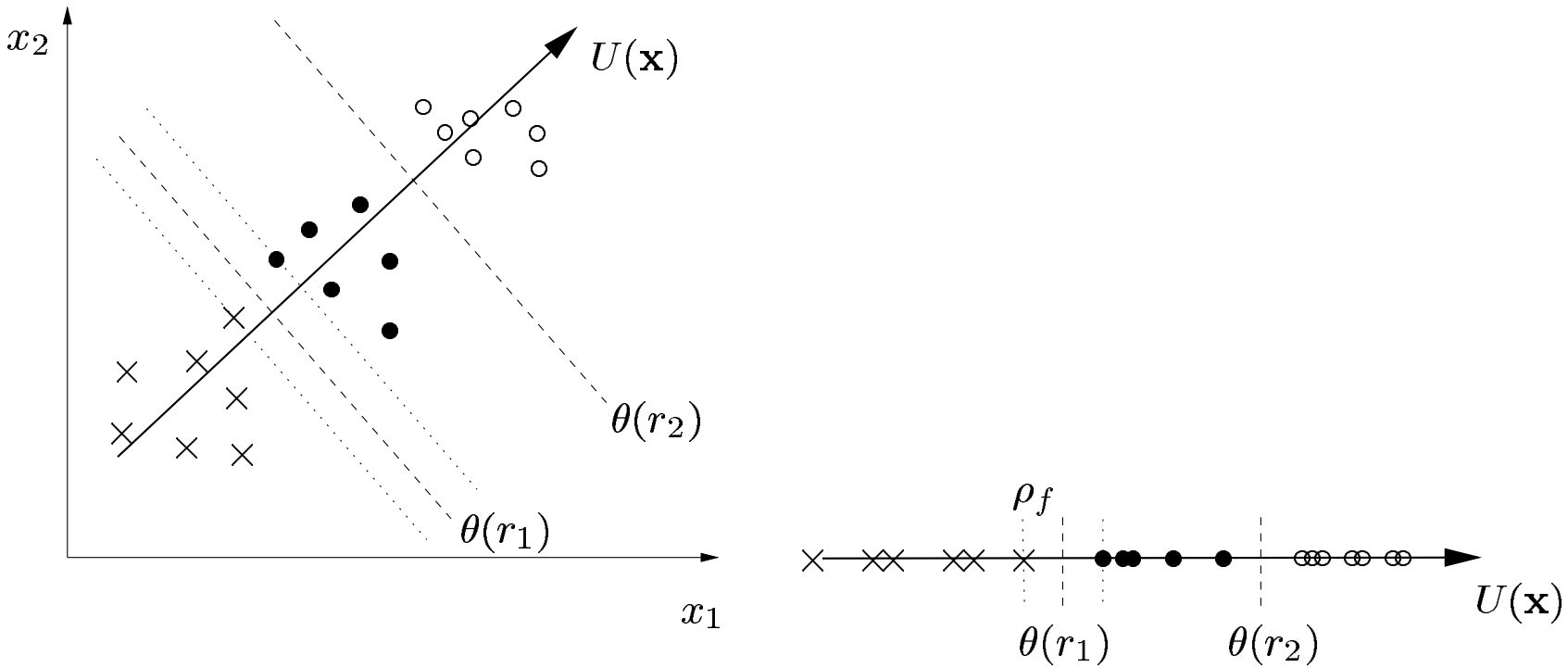
Class	b_1	b_2	b_3	b_4
Chinese	1	0	0	1
German	0	0	1	0
Argentine	1	1	1	0

- $(0, 0, 0, 1) \rightarrow$ German
- $(0, 1, 0, 1) \rightarrow$ Chinese

Bottom Line

- Understand what your algorithm is doing when you ask it to multi class
- Features and training imbalance matter more than ever
- Debugging is often easier if **you** binarize the problem

SVM Ranking


x_2
$U(\mathbf{x})$
$\theta(r_2)$
$\theta(r_1)$
x_1
ρ_f
$U(\mathbf{x})$
$\theta(r_1)$
$\theta(r_2)$

Real(-ish) Data

Sets of five movies ranked by users

```
# Big Lebowski, The
1 qid:375 1:0.04 2:0.01 3:1.1 4:0.0 5:1.0 6:0.0 7:0.0
# School of Rock, The
2 qid:375 1:0.06 2:-0.00 3:0.7 4:0.0 5:1.0 6:0.0 7:0.0
 # While You Were Sleeping
3 qid:375 1:0.03 2:-0.01 3:0.04 4:0.0 5:1.0 6:0.0 7:0.0
# Clockwise
4 qid:375 1:-0.01 2:-0.02 3:0.04 4:0.0 5:1.0 6:0.0 7:0.0
 # Enchanted April
5 qid:375 1:0.02 2:-0.02 3:0.7 4:0.0 5:0.0 6:0.0 7:1.0
```

1: Year of the movie ($\mu = 1987$, var=266)

Real(-ish) Data

Sets of five movies ranked by users

```
# Big Lebowski, The
1 qid:375 1:0.04 2:0.01 3:1.1 4:0.0 5:1.0 6:0.0 7:0.0
# School of Rock, The
2 qid:375 1:0.06 2:-0.00 3:0.7 4:0.0 5:1.0 6:0.0 7:0.0
 # While You Were Sleeping
3 qid:375 1:0.03 2:-0.01 3:0.04 4:0.0 5:1.0 6:0.0 7:0.0
# Clockwise
4 qid:375 1:-0.01 2:-0.02 3:0.04 4:0.0 5:1.0 6:0.0 7:0.0
 # Enchanted April
5 qid:375 1:0.02 2:-0.02 3:0.7 4:0.0 5:0.0 6:0.0 7:1.0
```

2: Length of the movie ($\mu = 108$, var=569)

Real(-ish) Data

Sets of five movies ranked by users

```
# Big Lebowski, The
1 qid:375 1:0.04 2:0.01 3:1.1 4:0.0 5:1.0 6:0.0 7:0.0
# School of Rock, The
2 qid:375 1:0.06 2:-0.00 3:0.7 4:0.0 5:1.0 6:0.0 7:0.0
 # While You Were Sleeping
3 qid:375 1:0.03 2:-0.01 3:0.04 4:0.0 5:1.0 6:0.0 7:0.0
# Clockwise
4 qid:375 1:-0.01 2:-0.02 3:0.04 4:0.0 5:1.0 6:0.0 7:0.0
 # Enchanted April
5 qid:375 1:0.02 2:-0.02 3:0.7 4:0.0 5:0.0 6:0.0 7:1.0
```

3: Rating ($\mu = 6.4$, var=1.4)

Real(-ish) Data

Sets of five movies ranked by users

```
# Big Lebowski, The
1 qid:375 1:0.04 2:0.01 3:1.1 4:0.0 5:1.0 6:0.0 7:0.0
# School of Rock, The
2 qid:375 1:0.06 2:-0.00 3:0.7 4:0.0 5:1.0 6:0.0 7:0.0
 # While You Were Sleeping
3 qid:375 1:0.03 2:-0.01 3:0.04 4:0.0 5:1.0 6:0.0 7:0.0
# Clockwise
4 qid:375 1:-0.01 2:-0.02 3:0.04 4:0.0 5:1.0 6:0.0 7:0.0
 # Enchanted April
5 qid:375 1:0.02 2:-0.02 3:0.7 4:0.0 5:0.0 6:0.0 7:1.0
```

4: Action (binary)

Real(-ish) Data

Sets of five movies ranked by users

```
# Big Lebowski, The
1 qid:375 1:0.04 2:0.01 3:1.1 4:0.0 5:1.0 6:0.0 7:0.0
# School of Rock, The
2 qid:375 1:0.06 2:-0.00 3:0.7 4:0.0 5:1.0 6:0.0 7:0.0
 # While You Were Sleeping
3 qid:375 1:0.03 2:-0.01 3:0.04 4:0.0 5:1.0 6:0.0 7:0.0
# Clockwise
4 qid:375 1:-0.01 2:-0.02 3:0.04 4:0.0 5:1.0 6:0.0 7:0.0
 # Enchanted April
5 qid:375 1:0.02 2:-0.02 3:0.7 4:0.0 5:0.0 6:0.0 7:1.0
```

5: Comedy (binary)

Real(-ish) Data

Sets of five movies ranked by users

```
# Big Lebowski, The
1 qid:375 1:0.04 2:0.01 3:1.1 4:0.0 5:1.0 6:0.0 7:0.0
# School of Rock, The
2 qid:375 1:0.06 2:-0.00 3:0.7 4:0.0 5:1.0 6:0.0 7:0.0
 # While You Were Sleeping
3 qid:375 1:0.03 2:-0.01 3:0.04 4:0.0 5:1.0 6:0.0 7:0.0
# Clockwise
4 qid:375 1:-0.01 2:-0.02 3:0.04 4:0.0 5:1.0 6:0.0 7:0.0
 # Enchanted April
5 qid:375 1:0.02 2:-0.02 3:0.7 4:0.0 5:0.0 6:0.0 7:1.0
```

6: Documentary (binary)

Real(-ish) Data

Sets of five movies ranked by users

```
# Big Lebowski, The
1 qid:375 1:0.04 2:0.01 3:1.1 4:0.0 5:1.0 6:0.0 7:0.0
# School of Rock, The
2 qid:375 1:0.06 2:-0.00 3:0.7 4:0.0 5:1.0 6:0.0 7:0.0
 # While You Were Sleeping
3 qid:375 1:0.03 2:-0.01 3:0.04 4:0.0 5:1.0 6:0.0 7:0.0
# Clockwise
4 qid:375 1:-0.01 2:-0.02 3:0.04 4:0.0 5:1.0 6:0.0 7:0.0
 # Enchanted April
5 qid:375 1:0.02 2:-0.02 3:0.7 4:0.0 5:0.0 6:0.0 7:1.0
```

7: Drama (binary)

Fitting an SVM

- SVM-RANK
- Five support vectors
- Weight vector

$$w = \langle 0.02, 0.03, -1.82, -2.30, -0.05, 1.73, 1.84 \rangle \tag{1}$$

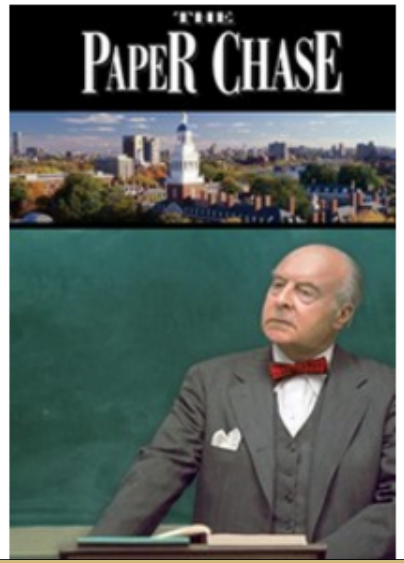
Fitting an SVM

- SVM-RANK
- Five support vectors
- Weight vector

$$w = \langle 0.02, 0.03, -1.82, -2.30, -0.05, 1.73, 1.84 \rangle \quad (1)$$

Target: older, shorter action movies with high ratings

Predictions



TOMATOMETER

All Critics | Top Critics

85%

Average Rating: 7.2/10
Reviews Counted: 26
Fresh: 22
Rotten: 4

Critics Consensus: No consensus yet.

AUDIENCE SCORE

71%
liked it

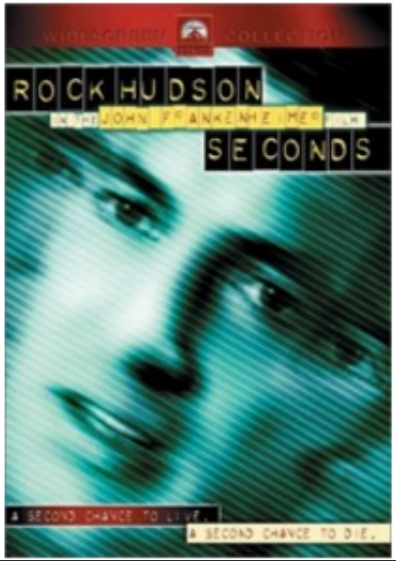
Average Rating: 3.5/5
User Ratings: 4,395

ADD YOUR RATING

WANT TO SEE | NOT INTERESTED

Add a Review (Optional)

Predictions



TOMATOMETER

All Critics | Top Critics

90%

Average Rating: 8.2/10
Reviews Counted: 31
Fresh: 28
Rotten: 3

Critics Consensus: Featuring dazzling, disorienting cinematography from the great James Wong Howe and a strong lead performance by Rock Hudson, Seconds is a compellingly paranoid take on the legend of Faust.

AUDIENCE SCORE

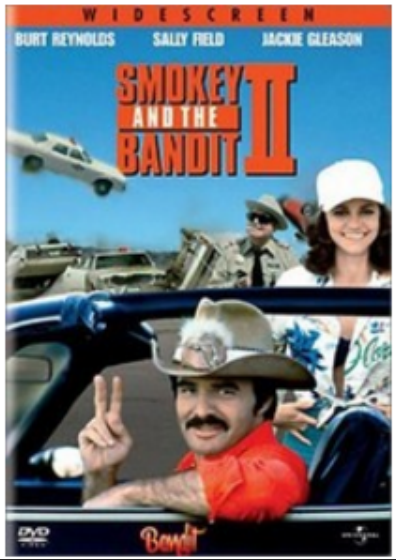
88%
liked it

Average Rating: 4/5
User Ratings: 4,246

ADD YOUR RATING

WANT TO SEE | NOT INTERESTED

Predictions



TOMATOMETER

All Critics | Top Critics

20%

Average Rating: 4/10
Reviews Counted: 5
Fresh: 1
Rotten: 4

Critics Consensus: No consensus yet.

AUDIENCE SCORE

45% liked it

Average Rating: 2.8/5
User Ratings: 33,546

ADD YOUR RATING

WANT TO SEE | NOT INTERESTED

Add a Review (Optional)

Predictions

TOMATOMETER | All Critics | Top Critics

52%

Average Rating: 5.6/10
Reviews Counted: 21
Fresh: 11
Rotten: 10

Critics Consensus: No consensus yet.

AUDIENCE SCORE

61% liked it

Average Rating: 3.2/5
User Ratings: 10,398

ADD YOUR RATING

WANT TO SEE | NOT INTERESTED

Add a Review (Optional)

Predictions

TOMATOMETER

All Critics | Top Critics

CERTIFIED FRESH Rotten Tomatoes **96%**

Average Rating: 8.5/10
Reviews Counted: 223
Fresh: 215
Rotten: 8

Critics Consensus: Charming, thoughtful, and often funny, Sideways is a decidedly mature road trip comedy full of excellent performances.

AUDIENCE SCORE

78% liked it

Average Rating: 3.4/5
User Ratings: 196,279

ADD YOUR RATING

+ WANT TO SEE | NOT INTERESTED

Add a Review (Optional)

Predictions

$$w = \langle 0.02, 0.03, -1.82, -2.30, -0.05, 1.73, 1.84 \rangle \tag{2}$$

```
# Paper Chase
1:-0.06 2:0.0 3:0.53 4:0.0 5:0.0 6:0.0 7:1.0
# Seconds
1:-0.08 2:-0.01 3:0.74 4:0.0 5:0.0 6:0.0 7:1.0
#Smokey and the Bandit II
1:-0.03 2:-0.02 3:-1.43 4:1.0 5:1.0 6:0.0 7:0.0
# CB4
1:0.02 2:-0.03 3:-0.73 4:0.0 5:1.0 6:0.0 7:0.0
#Sideways
1:0.06 2:0.03 3:1.09 4:0.0 5:1.0 6:0.0 7:1.0
```

- Paper Chase:

- Paper Chase: $-0.01 \cdot -0.06 + 0.07 \cdot 0.00 + -1.95 \cdot 0.53 + -2.28 \cdot 0.00 + -0.07 \cdot 0.00 + 1.57 \cdot 0.00 + 1.87 \cdot 1.00 = 0.84$

- Paper Chase: $-0.01 \cdot -0.06 + 0.07 \cdot 0.00 + -1.95 \cdot 0.53 + -2.28 \cdot 0.00 + -0.07 \cdot 0.00 + 1.57 \cdot 0.00 + 1.87 \cdot 1.00 = 0.84$
- Seconds:

- Paper Chase: $-0.01 \cdot -0.06 + 0.07 \cdot 0.00 + -1.95 \cdot 0.53 + -2.28 \cdot 0.00 + -0.07 \cdot 0.00 + 1.57 \cdot 0.00 + 1.87 \cdot 1.00 = 0.84$
- Seconds: $-0.01 \cdot -0.08 + 0.07 \cdot -0.01 + -1.95 \cdot 0.74 + -2.28 \cdot 0.00 + -0.07 \cdot 0.00 + 1.57 \cdot 0.00 + 1.87 \cdot 1.00 = 0.43$

- Paper Chase: $-0.01 \cdot -0.06 + 0.07 \cdot 0.00 + -1.95 \cdot 0.53 + -2.28 \cdot 0.00 + -0.07 \cdot 0.00 + 1.57 \cdot 0.00 + 1.87 \cdot 1.00 = 0.84$
- Seconds: $-0.01 \cdot -0.08 + 0.07 \cdot -0.01 + -1.95 \cdot 0.74 + -2.28 \cdot 0.00 + -0.07 \cdot 0.00 + 1.57 \cdot 0.00 + 1.87 \cdot 1.00 = 0.43$
- Smokey and the Bandit II:

- Paper Chase: $-0.01 \cdot -0.06 + 0.07 \cdot 0.00 + -1.95 \cdot 0.53 + -2.28 \cdot 0.00 + -0.07 \cdot 0.00 + 1.57 \cdot 0.00 + 1.87 \cdot 1.00 = 0.84$
- Seconds: $-0.01 \cdot -0.08 + 0.07 \cdot -0.01 + -1.95 \cdot 0.74 + -2.28 \cdot 0.00 + -0.07 \cdot 0.00 + 1.57 \cdot 0.00 + 1.87 \cdot 1.00 = 0.43$
- Smokey and the Bandit II: $-0.01 \cdot -0.03 + 0.07 \cdot -0.02 + -1.95 \cdot -1.43 + -2.28 \cdot 1.00 + -0.07 \cdot 1.00 + 1.57 \cdot 0.00 + 1.87 \cdot 0.00 = 0.44$

- Paper Chase: $-0.01 \cdot -0.06 + 0.07 \cdot 0.00 + -1.95 \cdot 0.53 + -2.28 \cdot 0.00 + -0.07 \cdot 0.00 + 1.57 \cdot 0.00 + 1.87 \cdot 1.00 = 0.84$
- Seconds: $-0.01 \cdot -0.08 + 0.07 \cdot -0.01 + -1.95 \cdot 0.74 + -2.28 \cdot 0.00 + -0.07 \cdot 0.00 + 1.57 \cdot 0.00 + 1.87 \cdot 1.00 = 0.43$
- Smokey and the Bandit II: $-0.01 \cdot -0.03 + 0.07 \cdot -0.02 + -1.95 \cdot -1.43 + -2.28 \cdot 1.00 + -0.07 \cdot 1.00 + 1.57 \cdot 0.00 + 1.87 \cdot 0.00 = 0.44$
- CB4:

- Paper Chase: $-0.01 \cdot -0.06 + 0.07 \cdot 0.00 + -1.95 \cdot 0.53 + -2.28 \cdot 0.00 + -0.07 \cdot 0.00 + 1.57 \cdot 0.00 + 1.87 \cdot 1.00 = 0.84$
- Seconds: $-0.01 \cdot -0.08 + 0.07 \cdot -0.01 + -1.95 \cdot 0.74 + -2.28 \cdot 0.00 + -0.07 \cdot 0.00 + 1.57 \cdot 0.00 + 1.87 \cdot 1.00 = 0.43$
- Smokey and the Bandit II: $-0.01 \cdot -0.03 + 0.07 \cdot -0.02 + -1.95 \cdot -1.43 + -2.28 \cdot 1.00 + -0.07 \cdot 1.00 + 1.57 \cdot 0.00 + 1.87 \cdot 0.00 = 0.44$
- CB4: $0.01 \cdot 0.02 + 0.07 \cdot -0.03 + -1.95 \cdot -0.73 + -2.28 \cdot 0.00 + -0.07 \cdot 1.00 + 1.57 \cdot 0.00 + 1.87 \cdot 0.00 = 1.35$

- Paper Chase: $-0.01 \cdot -0.06 + 0.07 \cdot 0.00 + -1.95 \cdot 0.53 + -2.28 \cdot 0.00 + -0.07 \cdot 0.00 + 1.57 \cdot 0.00 + 1.87 \cdot 1.00 = 0.84$
- Seconds: $-0.01 \cdot -0.08 + 0.07 \cdot -0.01 + -1.95 \cdot 0.74 + -2.28 \cdot 0.00 + -0.07 \cdot 0.00 + 1.57 \cdot 0.00 + 1.87 \cdot 1.00 = 0.43$
- Smokey and the Bandit II: $-0.01 \cdot -0.03 + 0.07 \cdot -0.02 + -1.95 \cdot -1.43 + -2.28 \cdot 1.00 + -0.07 \cdot 1.00 + 1.57 \cdot 0.00 + 1.87 \cdot 0.00 = 0.44$
- CB4: $0.01 \cdot 0.02 + 0.07 \cdot -0.03 + -1.95 \cdot -0.73 + -2.28 \cdot 0.00 + -0.07 \cdot 1.00 + 1.57 \cdot 0.00 + 1.87 \cdot 0.00 = 1.35$
- Sideways: $-0.01 \cdot 0.06 + 0.07 \cdot 0.03 + -1.95 \cdot 1.09 + -2.28 \cdot 0.00 + -0.07 \cdot 1.00 + 1.57 \cdot 0.00 + 1.87 \cdot 1.00 = -0.32$

- Paper Chase: $-0.01 \cdot -0.06 + 0.07 \cdot 0.00 + -1.95 \cdot 0.53 + -2.28 \cdot 0.00 + -0.07 \cdot 0.00 + 1.57 \cdot 0.00 + 1.87 \cdot 1.00 = 0.84$
- Seconds: $-0.01 \cdot -0.08 + 0.07 \cdot -0.01 + -1.95 \cdot 0.74 + -2.28 \cdot 0.00 + -0.07 \cdot 0.00 + 1.57 \cdot 0.00 + 1.87 \cdot 1.00 = 0.43$
- Smokey and the Bandit II: $-0.01 \cdot -0.03 + 0.07 \cdot -0.02 + -1.95 \cdot -1.43 + -2.28 \cdot 1.00 + -0.07 \cdot 1.00 + 1.57 \cdot 0.00 + 1.87 \cdot 0.00 = 0.44$
- CB4: $0.01 \cdot 0.02 + 0.07 \cdot -0.03 + -1.95 \cdot -0.73 + -2.28 \cdot 0.00 + -0.07 \cdot 1.00 + 1.57 \cdot 0.00 + 1.87 \cdot 0.00 = 1.35$
- Sideways: $-0.01 \cdot 0.06 + 0.07 \cdot 0.03 + -1.95 \cdot 1.09 + -2.28 \cdot 0.00 + -0.07 \cdot 1.00 + 1.57 \cdot 0.00 + 1.87 \cdot 1.00 = -0.32$

What's the predicted ranking?

Ranking

Predicted Rank

1. Sideways
2. Seconds
3. Smokey and the Bandit II
4. The Paper Chase
5. CB4

Ranking

Predicted Rank	True Rank
1. Sideways	1. Sideways
2. Seconds	2. Smokey and the Bandit II
3. Smokey and the Bandit II	3. Seconds
4. The Paper Chase	4. The Paper Chase
5. CB4	5. CB4

Ranking

Predicted Rank	True Rank
1. Sideways	1. Sideways
2. Seconds	2. Smokey and the Bandit II
3. Smokey and the Bandit II	3. Seconds
4. The Paper Chase	4. The Paper Chase
5. CB4	5. CB4

How many errors is this?

Ranking

Predicted Rank	True Rank
1. Sideways	1. Sideways
2. Seconds	2. Smokey and the Bandit II
3. Smokey and the Bandit II	3. Seconds
4. The Paper Chase	4. The Paper Chase
5. CB4	5. CB4

How many errors is this? S&B 2 $>$ Seconds

Ranking to Regression

- Using SVMs to predict a value
- Ranking that value
- What if we care about actual value and not just relative order?
- Regression!