

Data Mining (Week 1)

MSc Data Mining

Topic 05 : Classification

Foundation

~~Data Handling~~

Exploratory
Analysis

Part 03 : Naive Bayes

Data Modelling
Fundamentals

Data Modelling
Advanced

Rule Based

Association Rules

Recommender Systems

Unsupervised

Clustering

Spring Semester, 2022

Supervised

Regression

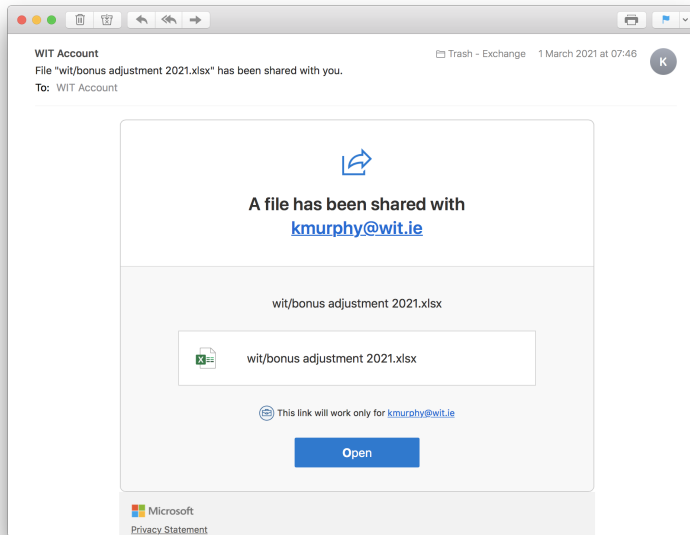
Basic Classification

Outline

- Motivation behind naïve Bayes
- Sample application
- Summary of properties

Spam Filtering

Reality



Simplified Problem

Assume that we have the following set of messages previously classified as **spam** or **ham**.

Message	Class
"send us your password"	spam
"send us your review"	ham
"password review"	ham
"review us"	spam
"send your password"	spam
"send us your account"	spam

We are interested in classifying the following new message as **spam** or **ham**:

Message	Class
"review us now"	?

Count occurrences ... compute probabilities

Message	Class	Word	# in spam	# in ham	$\Pr(\bullet \text{spam})$	$\Pr(\bullet \text{ham})$
“send us your password”	spam	account	1	0	1/4	0/2
“send us your review”	ham	password	2	1	2/4	1/2
“password review”	ham	review	1	2	1/4	2/2
“review us”	spam	send	3	1	3/4	1/2
“send your password”	spam	us	3	1	3/4	1/2
“send us your account”	spam	your	3	1	3/4	1/2

Class	Count	Probability
spam	4	$\Pr(\text{spam}) = 4/6$
ham	2	$\Pr(\text{ham}) = 2/6$

prior probabilities

What we have

The probability that a message contains, say the word **review**, among our message classed as **spam**, i.e.,

$$\Pr(\text{review}|\text{spam}) = 1/4$$

What we want

The probability that a message is **spam** given that it contains, say the word **review**, i.e.,

$$\Pr(\text{spam}|\text{review}) = ?$$

posterior probabilities

Aside: Probability Laws

Bayes Rule

$$\begin{array}{c}
 \text{conditional} \times \text{marginal} \\
 \underbrace{\Pr(A|B) \Pr(B)} = \underbrace{\Pr(A \text{ AND } B)}_{\text{joint}} = \underbrace{\Pr(B|A) \Pr(A)}_{\text{conditional} \times \text{marginal}}
 \end{array}$$

$$\Pr(A|B) \Pr(B) = \Pr(B|A) \Pr(A) \longrightarrow \Pr(B|A) = \frac{\Pr(A|B) \Pr(B)}{\Pr(A)}$$

Complementary Rule

$$\Pr(\text{NOT } A) = 1 - \Pr(A)$$

Independent Events

If events A and B are independent, then $\Pr(A \text{ AND } B) = \Pr(A) \Pr(B)$

Law of total of probability

Let A be any event. Let $\{B_1, B_2, B_3, \dots, B_n\}$ be a set of events, exactly one of which must occur, then

$$\Pr(A) = \Pr(A|B_1) \Pr(B_1) + \Pr(A|B_2) \Pr(B_2) + \dots + \Pr(A|B_n) \Pr(B_n)$$

Classifying a Message Using a Single Word

Applying the law of total probability, with $A = \text{review}$, and $B_1 = \text{spam}$ and $B_2 = \text{ham}$, then we have

$$\Pr(\text{review}) = \Pr(\text{review}|\text{spam}) \Pr(\text{spam}) + \Pr(\text{review}|\text{ham}) \Pr(\text{ham}) = \frac{1}{4} \cdot \frac{4}{6} + \frac{2}{2} \cdot \frac{2}{6} = \frac{3}{6}$$

Now we can apply Bayes rule, with $A = \text{review}$ and $B = \text{spam}$ we have

$$\Pr(\text{spam}|\text{review}) = \frac{\Pr(\text{review}|\text{spam}) \Pr(\text{spam})}{\Pr(\text{review})} = \frac{\frac{1}{4} \cdot \frac{4}{6}}{\frac{3}{6}} = \frac{1}{3} = 33.3\%$$

And we can apply Bayes rule, with $A = \text{review}$ and $B = \text{ham}$ to get

$$\Pr(\text{ham}|\text{review}) = \frac{\Pr(\text{review}|\text{ham}) \Pr(\text{ham})}{\Pr(\text{review})} = \frac{\frac{2}{2} \cdot \frac{2}{6}}{\frac{3}{6}} = \frac{2}{3} = 66.7\%$$

So on receiving a message containing the word **review** we can classify it a **spam** with probability 33.3% and **ham** with probability 66.7%.

Classifying a Message Using Multiple Words

I

- We can now (hopefully) do this for each word in our test message (“review us now”), but what about classifying using multiple words? ... here comes the naïve bit ...

Naïve Bayes assumes that presence of each word are independent events

Recall: independent events means can multiply to get joint probabilities.

- We are interested in classifying the message

review us now

- We don't have data on the word **now** so only looking at messages containing **review** and **us** and not containing **account**, **password**, **send**, or **your**.

$$\Pr(\{\text{review}, \text{us}\} | \text{spam}) = \underbrace{\left(1 - \frac{1}{4}\right)}_{\text{account}} \underbrace{\left(1 - \frac{2}{4}\right)}_{\text{password}} \underbrace{\left(\frac{1}{4}\right)}_{\text{review}} \underbrace{\left(1 - \frac{3}{4}\right)}_{\text{send}} \underbrace{\left(\frac{3}{4}\right)}_{\text{us}} \underbrace{\left(1 - \frac{3}{4}\right)}_{\text{your}} = 0.0044$$

and

$$\Pr(\{\text{review}, \text{us}\} | \text{ham}) = \underbrace{\left(1 - \frac{0}{2}\right)}_{\text{account}} \underbrace{\left(1 - \frac{1}{2}\right)}_{\text{password}} \underbrace{\left(\frac{2}{2}\right)}_{\text{review}} \underbrace{\left(1 - \frac{1}{2}\right)}_{\text{send}} \underbrace{\left(\frac{1}{2}\right)}_{\text{us}} \underbrace{\left(1 - \frac{1}{2}\right)}_{\text{your}} = 0.0625$$

Classifying a Message Using Multiple Words

Now we can apply the law of total probability as before

$$\begin{aligned}\Pr(\{\text{review}, \text{us}\}) &= \Pr(\{\text{review}, \text{us}\}|\text{spam}) \Pr(\text{spam}) + \Pr(\{\text{review}, \text{us}\}|\text{ham}) \Pr(\text{ham}) \\ &= 0.0044 \left(\frac{4}{6}\right) + 0.0625 \left(\frac{2}{6}\right) = 0.0237\end{aligned}$$

And finally Bayes rule

$$\Pr(\text{spam}|\{\text{review}, \text{us}\}) = \frac{\Pr(\{\text{review}, \text{us}\}|\text{spam}) \Pr(\text{spam})}{\Pr(\{\text{review}, \text{us}\})} = \frac{0.0044 \left(\frac{4}{6}\right)}{0.0237} = 0.123$$

and

$$\Pr(\text{ham}|\{\text{review}, \text{us}\}) = \frac{\Pr(\{\text{review}, \text{us}\}|\text{ham}) \Pr(\text{ham})}{\Pr(\{\text{review}, \text{us}\})} = \frac{0.0625 \left(\frac{2}{6}\right)}{0.0237} = 0.877$$

Hence, the probability that the message is **spam** is 12.3%, and **ham** is 87.7%.

Classifying a Message Using Multiple Words

III

A final comment:

- The classifier picks the class with the largest probability, so in this case

$$\Pr(\text{spam}|\{\text{review}, \text{us}\}) = \frac{\Pr(\{\text{review}, \text{us}\}|\text{spam}) \Pr(\text{spam})}{\Pr(\{\text{review}, \text{us}\})} = \frac{0.0044 \left(\frac{4}{6}\right)}{0.0237} = 0.123$$

and

$$\Pr(\text{ham}|\{\text{review}, \text{us}\}) = \frac{\Pr(\{\text{review}, \text{us}\}|\text{ham}) \Pr(\text{ham})}{\Pr(\{\text{review}, \text{us}\})} = \frac{0.0625 \left(\frac{2}{6}\right)}{0.0237} = 0.877$$

- We can skip the computation of $\Pr(\{\text{review}, \text{us}\})$ involving the law of total probability step since that is a common denominator in above formula, And instead look are relative likelihood

$$\frac{\Pr(\text{spam}|\{\text{review}, \text{us}\})}{\Pr(\text{ham}|\{\text{review}, \text{us}\})} = \frac{\Pr(\{\text{review}, \text{us}\}|\text{spam}) \Pr(\text{spam})}{\Pr(\{\text{review}, \text{us}\}|\text{ham}) \Pr(\text{ham})} \implies \begin{cases} > 1 & \text{pick spam} \\ < 1 & \text{pick ham} \end{cases}$$

Trump's Claims — Dataset

The Washington Post has released a curated list of Donald Trump's false claims while in office. The claim's have been categorised and cross linked. We will see how well a naïve Bayes classifier will do on this dataset.

```
df = pd.read_csv("wapo_trumpclaims_export-012021.csv.gz")
print(df.shape)
df.head(5)
```

```
(30573, 9)
```

	id	location	claim	analysis	pinocchios	category	repeated_ids	repeated_count	date
0	31608.0	Remarks	"We also got tax cuts, the largest tax cut and...	This is Trump's second favorite falsehood, and...	4.0	Taxes	31608, 31581, 31305, 31183, 31530, 30920, 3085...	296	01/20/2021
1	31609.0	Remarks	"We just got seventy five million votes. And t...	When the counting was finished, Trump had rece...	NaN	Election	31609, 31292, 31155, 31016, 31082, 30992, 3156...	19	01/20/2021
2	31610.0	Remarks	"One of the things we're very, very proud of i...	Contrary to his boasts, Trump did not achieve ...	NaN	Miscellaneous	31610, 31598, 31187, 30918, 30374, 29845, 2922...	84	01/20/2021
3	31611.0	Remarks	"Our first lady has been a woman of great grac...	In reality, Melania Trump leaves the White Hou...	NaN	Miscellaneous	NaN	0	01/20/2021
4	31612.0	Remarks	"That's why [regulation cuts] we have such goo...	Leaving aside Trump's claim about the impact o...	NaN	Jobs	NaN	0	01/20/2021

Trump's Claims — Dataset

2 `df.category.value_counts(dropna=False)`

Immigration	3225
Foreign policy	3165
Election	3037
Miscellaneous	2767
Coronavirus	2521
Trade	2513
Economy	2475
Russia	1838
Jobs	1732
Health care	1629
Ukraine probe	1377
Environment	1065
Biographical record	963
Taxes	857
Crime	852
NaN	169
Guns	165
Education	151
Terrorism	72

Name: category, dtype: int64

General observations ...

- Some classes are relatively rare.
- Some classifications would appear to overlap Russia ↔ Election.
- There are 169 unclassified observations, missing values ... **these need** to be removed/encoded before model building.
- Total of 30,573 false claims! He was busy between golf sessions ... sorry, I could not resist.

Trump's Claims — Data Preprocessing

II

We perform our standard train-test split

```
3 from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(
    df.claim, df.category, test_size=0.2, random_state=42)
```

Next we need to reproduce the word extraction and counts that we did for the simple example earlier:

```
4 from sklearn.feature_extraction.text import CountVectorizer
count_vectorizer = CountVectorizer()

X_train_counts = count_vectorizer.fit_transform(X_train)
X_test_counts = count_vectorizer.transform(X_test)
```

Use the train dataset to determine parameters for the operation (**fit**). Then, apply (**transform**) to both train and test

Notice the function calls `fit_transform` on `X_train` and `transform` on `X_test`. The function `fit_transform` is two separate steps: `fit` and `transform` and could be written as

```
5 count_vectorizer.fit(X_train)
X_train_counts = count_vectorizer.transform(X_train)
```

or

```
6 X_train_counts = count_vectorizer.fit(X_train).transform(X_train)
```

How many features?

```
print(count_vectorizer.get_feature_names()[:1000])
```

0th, 1st, 2nd, 3rd, 4th, 5th, 6th, 7th, 8th, 9th, 10th, 11th, 12th, 13th, 14th, 15th, 16th, 17th, 18th, 19th, 20th, 21st, 22nd, 23rd, 24th, 25th, 26th, 27th, 28th, 29th, 30th, 31st, 32nd, 33rd, 34th, 35th, 36th, 37th, 38th, 39th, 40th, 41st, 42nd, 43rd, 44th, 45th, 46th, 47th, 48th, 49th, 50th, 51st, 52nd, 53rd, 54th, 55th, 56th, 57th, 58th, 59th, 60th, 61st, 62nd, 63rd, 64th, 65th, 66th, 67th, 68th, 69th, 70th, 71st, 72nd, 73rd, 74th, 75th, 76th, 77th, 78th, 79th, 80th, 81st, 82nd, 83rd, 84th, 85th, 86th, 87th, 88th, 89th, 90th, 91st, 92nd, 93rd, 94th, 95th, 96th, 97th, 98th, 99th, 100th, 101st, 102nd, 103rd, 104th, 105th, 106th, 107th, 108th, 109th, 110th, 111th, 112th, 113th, 114th, 115th, 116th, 117th, 118th, 119th, 120th, 121st, 122nd, 123rd, 124th, 125th, 126th, 127th, 128th, 129th, 130th, 131st, 132nd, 133rd, 134th, 135th, 136th, 137th, 138th, 139th, 140th, 141st, 142nd, 143rd, 144th, 145th, 146th, 147th, 148th, 149th, 150th, 151st, 152nd, 153rd, 154th, 155th, 156th, 157th, 158th, 159th, 160th, 161st, 162nd, 163rd, 164th, 165th, 166th, 167th, 168th, 169th, 170th, 171st, 172nd, 173rd, 174th, 175th, 176th, 177th, 178th, 179th, 180th, 181st, 182nd, 183rd, 184th, 185th, 186th, 187th, 188th, 189th, 190th, 191st, 192nd, 193rd, 194th, 195th, 196th, 197th, 198th, 199th, 200th, 201st, 202nd, 203rd, 204th, 205th, 206th, 207th, 208th, 209th, 210th, 211st, 212nd, 213th, 214th, 215th, 216th, 217th, 218th, 219th, 220th, 221st, 222nd, 223rd, 224th, 225th, 226th, 227th, 228th, 229th, 230th, 231st, 232nd, 233rd, 234th, 235th, 236th, 237th, 238th, 239th, 240th, 241st, 242nd, 243rd, 244th, 245th, 246th, 247th, 248th, 249th, 250th, 251st, 252nd, 253rd, 254th, 255th, 256th, 257th, 258th, 259th, 260th, 261st, 262nd, 263rd, 264th, 265th, 266th, 267th, 268th, 269th, 270th, 271st, 272nd, 273rd, 274th, 275th, 276th, 277th, 278th, 279th, 280th, 281st, 282nd, 283rd, 284th, 285th, 286th, 287th, 288th, 289th, 290th, 291st, 292nd, 293rd, 294th, 295th, 296th, 297th, 298th, 299th, 300th, 301st, 302nd, 303rd, 304th, 305th, 306th, 307th, 308th, 309th, 310th, 311st, 312nd, 313th, 314th, 315th, 316th, 317th, 318th, 319th, 320th, 321st, 322nd, 323rd, 324th, 325th, 326th, 327th, 328th, 329th, 330th, 331st, 332nd, 333rd, 334th, 335th, 336th, 337th, 338th, 339th, 340th, 341st, 342nd, 343rd, 344th, 345th, 346th, 347th, 348th, 349th, 350th, 351st, 352nd, 353rd, 354th, 355th, 356th, 357th, 358th, 359th, 360th, 361st, 362nd, 363rd, 364th, 365th, 366th, 367th, 368th, 369th, 370th, 371st, 372nd, 373rd, 374th, 375th, 376th, 377th, 378th, 379th, 380th, 381st, 382nd, 383rd, 384th, 385th, 386th, 387th, 388th, 389th, 390th, 391st, 392nd, 393rd, 394th, 395th, 396th, 397th, 398th, 399th, 400th, 401st, 402nd, 403rd, 404th, 405th, 406th, 407th, 408th, 409th, 410th, 411st, 412nd, 413th, 414th, 415th, 416th, 417th, 418th, 419th, 420th, 421st, 422nd, 423rd, 424th, 425th, 426th, 427th, 428th, 429th, 430th, 431st, 432nd, 433rd, 434th, 435th, 436th, 437th, 438th, 439th, 440th, 441st, 442nd, 443rd, 444th, 445th, 446th, 447th, 448th, 449th, 450th, 451st, 452nd, 453rd, 454th, 455th, 456th, 457th, 458th, 459th, 460th, 461st, 462nd, 463rd, 464th, 465th, 466th, 467th, 468th, 469th, 470th, 471st, 472nd, 473rd, 474th, 475th, 476th, 477th, 478th, 479th, 480th, 481st, 482nd, 483rd, 484th, 485th, 486th, 487th, 488th, 489th, 490th, 491st, 492nd, 493rd, 494th, 495th, 496th, 497th, 498th, 499th, 500th, 501st, 502nd, 503rd, 504th, 505th, 506th, 507th, 508th, 509th, 510th, 511st, 512nd, 513th, 514th, 515th, 516th, 517th, 518th, 519th, 520th, 521st, 522nd, 523rd, 524th, 525th, 526th, 527th, 528th, 529th, 530th, 531st, 532nd, 533rd, 534th, 535th, 536th, 537th, 538th, 539th, 540th, 541st, 542nd, 543rd, 544th, 545th, 546th, 547th, 548th, 549th, 550th, 551st, 552nd, 553rd, 554th, 555th, 556th, 557th, 558th, 559th, 560th, 561st, 562nd, 563rd, 564th, 565th, 566th, 567th, 568th, 569th, 570th, 571st, 572nd, 573rd, 574th, 575th, 576th, 577th, 578th, 579th, 580th, 581st, 582nd, 583rd, 584th, 585th, 586th, 587th, 588th, 589th, 590th, 591st, 592nd, 593rd, 594th, 595th, 596th, 597th, 598th, 599th, 600th, 601st, 602nd, 603rd, 604th, 605th, 606th, 607th, 608th, 609th, 610th, 611st, 612nd, 613th, 614th, 615th, 616th, 617th, 618th, 619th, 620th, 621st, 622nd, 623rd, 624th, 625th, 626th, 627th, 628th, 629th, 630th, 631st, 632nd, 633rd, 634th, 635th, 636th, 637th, 638th, 639th, 640th, 641st, 642nd, 643rd, 644th, 645th, 646th, 647th, 648th, 649th, 650th, 651st, 652nd, 653rd, 654th, 655th, 656th, 657th, 658th, 659th, 660th, 661st, 662nd, 663rd, 664th, 665th, 666th, 667th, 668th, 669th, 670th, 671st, 672nd, 673rd, 674th, 675th, 676th, 677th, 678th, 679th, 680th, 681st, 682nd, 683rd, 684th, 685th, 686th, 687th, 688th, 689th, 690th, 691st, 692nd, 693rd, 694th, 695th, 696th, 697th, 698th, 699th, 700th

Trump's Claims — Model Fit, Prediction and Evaluation

I

As before* we import our model and create an instance of it

```
10 from sklearn.naive_bayes import MultinomialNB  
model = MultinomialNB()
```

Fit our model using the train dataset

```
11 model.fit(X_train_counts, y_train)
```

Generate predictions using the test dataset

```
12 y_pred = model.predict(X_test_counts)
```

How well did it do? ... using accuracy score we have

```
13 from sklearn.metrics import accuracy_score  
accuracy_score(y_test, y_pred)
```

```
0.7763525735898701
```

This is not bad ... there are lots of class levels and some overlap between classes.

*This is the beauty of sklearn — consistent interface to all models, pre-/post- processing steps.

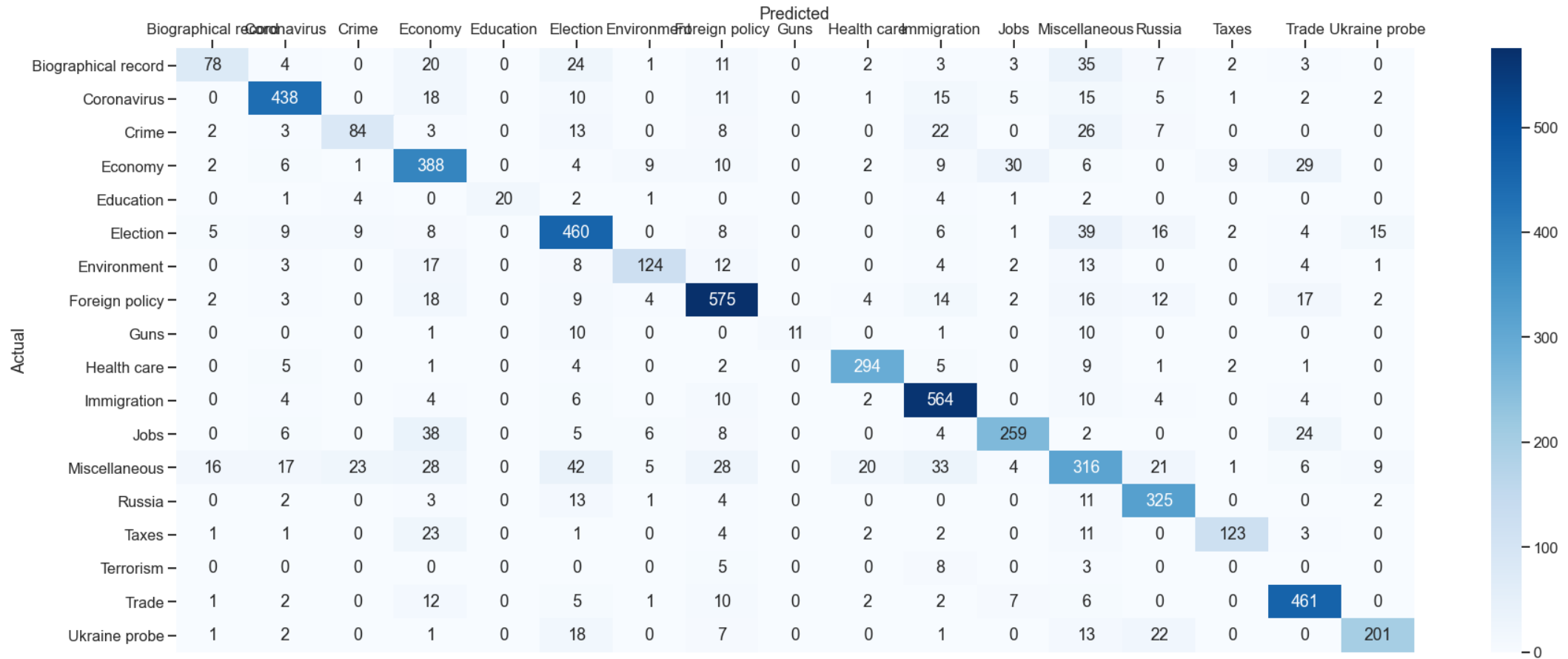
Trump's Claims —Model Fit, Prediction and Evaluation

II

Predicted	Biographical record	Coronavirus	Crime	Economy	Education	Election	Environment	Foreign policy	Guns	Health care	Immigration	Jobs	Miscellaneous	Russia	Tax
Actual															
Biographical record	78	4	0	20	0	24	1	11	0	2	3	3	35	7	2
Coronavirus	0	438	0	18	0	10	0	11	0	1	15	5	15	5	1
Crime	2	3	84	3	0	13	0	8	0	0	22	0	26	7	0
Economy	2	6	1	388	0	4	9	10	0	2	9	30	6	0	9
Education	0	1	4	0	20	2	1	0	0	0	4	1	2	0	0
Election	5	9	9	8	0	460	0	8	0	0	6	1	39	16	2
Environment	0	3	0	17	0	8	124	12	0	0	4	2	13	0	0
Foreign policy	2	3	0	18	0	9	4	575	0	4	14	2	16	12	0
Guns	0	0	0	1	0	10	0	0	11	0	1	0	10	0	0
Health care	0	5	0	1	0	4	0	2	0	294	5	0	9	1	2
Immigration	0	4	0	4	0	6	0	10	0	2	564	0	10	4	0
Jobs	0	6	0	38	0	5	6	8	0	0	4	259	2	0	0
Miscellaneous	16	17	23	28	0	42	5	28	0	20	33	4	316	21	1
Russia	0	2	0	3	0	13	1	4	0	0	0	0	11	325	0
Taxes	1	1	0	23	0	1	0	4	0	2	2	0	11	0	123
Terrorism	0	0	0	0	0	0	0	5	0	0	8	0	3	0	0
Trade	1	2	0	12	0	5	1	10	0	2	2	7	6	0	0
Ukraine probe	1	2	0	1	0	18	0	7	0	0	1	0	13	22	0

Trump's Claims — Model Fit, Prediction and Evaluation

III



- Look at the largest, off main-diagonal entries.

Naïve Bayes Classifier — Review

When to Consider

- Assumption of independence holds
- Categorical features.
- Spam filtering, Sentiment Analysis, and Recommendation Systems (with collaborative filtering).
- Variants exist for numeric (Gaussian-), binary (Bernoulli-) and multi-class (multinomial-) featured Naïve Bayes.

Advantages

- It is easy and fast to predict class. It also perform well in multi-class prediction.
- When assumption of independence holds, performs better compare to other models like logistic regression and needs less training data.

Disadvantages

- Ignores feature relationships
- If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as “Zero Frequency”. To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is Laplace estimation.
- If continuous features, then assumes normality conditions — often too restrictive.

Resources

- **A Gentle Introduction to Bayes Theorem for Machine Learning**

<https://machinelearningmastery.com/bayes-theorem-for-machine-learning/>

This is well worth a read over a cup of coffee. The author, Jason Brownlee, is worth following.