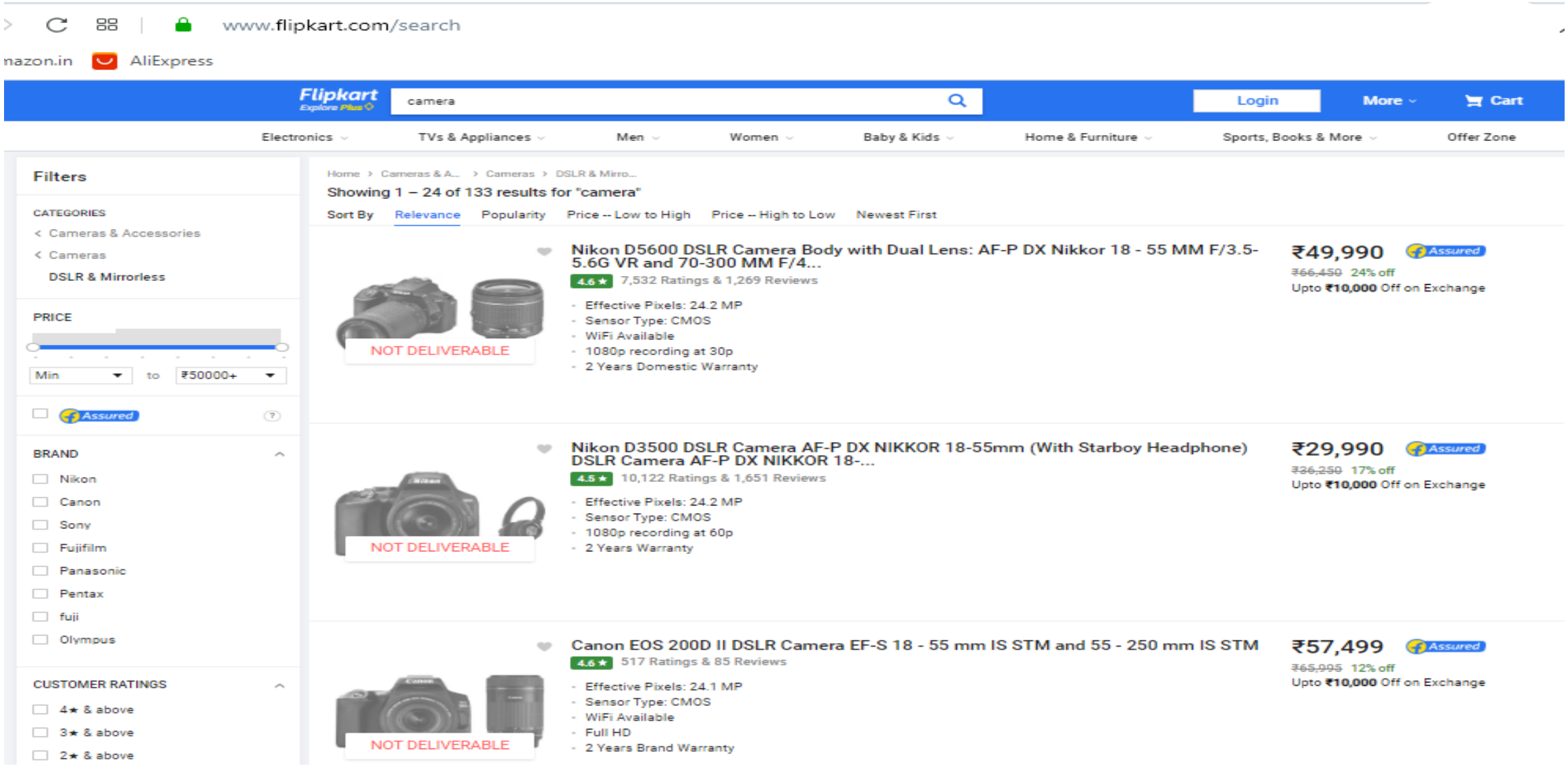
Retrieval and Feature Selection for Product Recommendation

Machine Learning Unit 7

Product Recommendation



Product Recommendation

- Based on cumulative ratings by all customers
- Based on similarity with product that user is currently viewing
  - based on customer statistics
  - based on product features
- Based on user's personal history
  - which products is the user likely to like???

Product Recommendation as Retrieval

- Input: a product that user is currently viewing (query product)
- Output: a set of other products which are “similar” to the input

- But how to evaluate the output?
- What if only a small number of products can be displayed, while many more are classified as “similar”?

Evaluation of Retrieval

- Output: set of products “A” which the algorithm considers to be similar to the query product
- Ground truth: set of products “B” in the database which experts consider to be similar to the query product

- Precision = $|A \cap B| / |A|$ (what fraction of the retrieved products are part of the ground truth?)
- Recall = $|A \cap B| / |B|$ (what fraction of the ground truth are retrieved?)
- F-score = harmonic mean of Precision and Recall

Bias of classification

- What if almost all labelled samples are negative?
- A database has millions of products, only a few are relevant to the query!
- Result: the classifier will be biased!
- Almost all test cases will be classified as negative!
- Overall accuracy: high, but positive test cases may be wrongly classified!
- To understand the bias, we need to measure classification accuracy for different classes separately!

TPR and FPR

- Consider one class as “positive”, other as “negative”
- A product considered “similar” to the query may be considered positive
- True Positive Rate (TPR): Out of all positive samples, what fraction is correctly predicted as “positive” (same as recall)
- False Positive Rate (FPR): Out of all negative samples what fraction is wrongly predicted as “positive”,
- Good classifier: high TPR, low FPR
- Biased classifier:

    i) Biased towards positive: high TPR, high FPR

    ii) Biased towards negative: low TPR, low FPR

Ranking of Results

- Classification: is a particular product “relevant” to the query or not?
- But we may need to measure “how relevant?” also
- Each product that is marked as “relevant”, should also get a similarity score!
- They can be sorted by the similarity score, and a small number selected

Ranking of Results

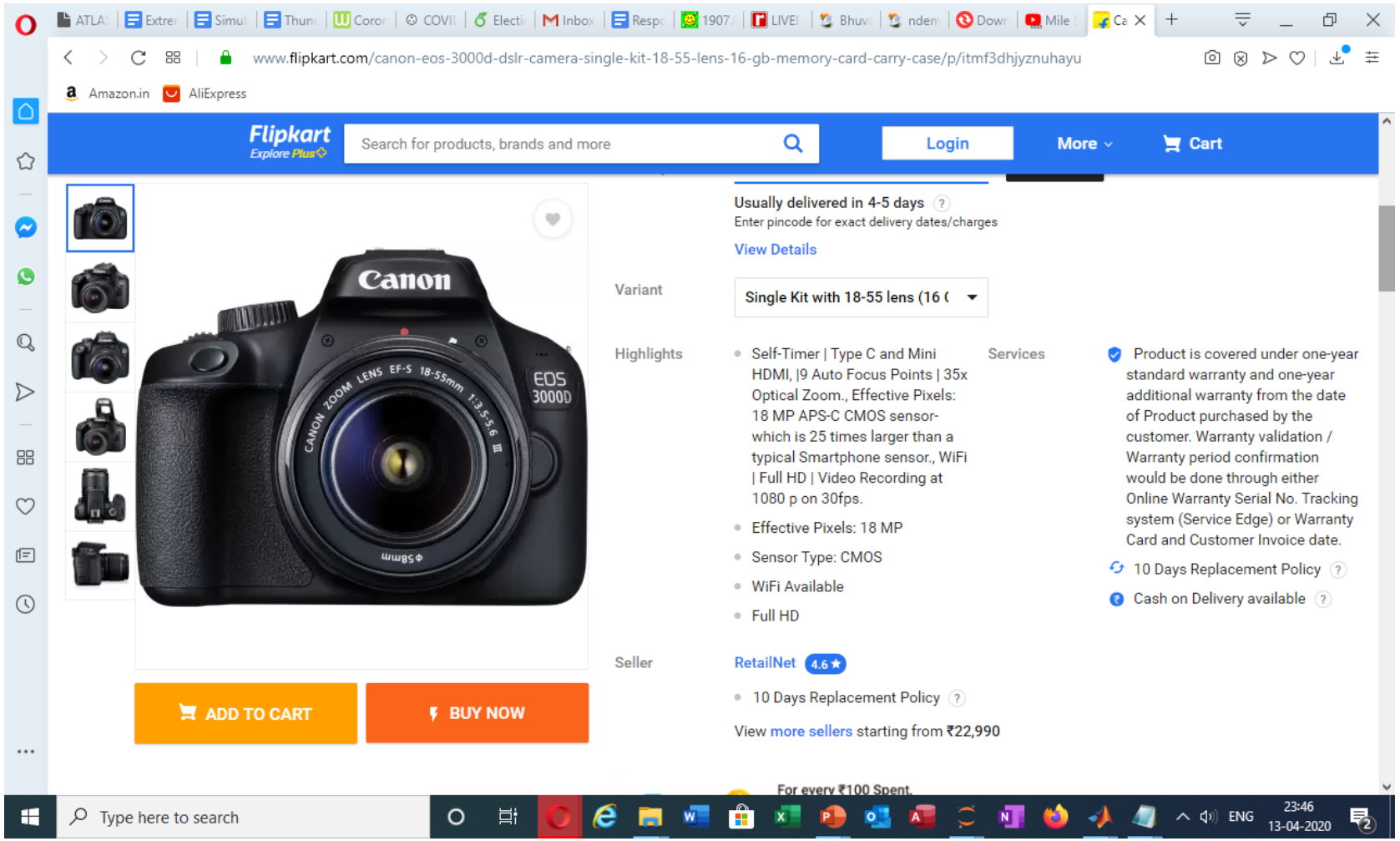
- How to get those scores?
- Consider some products identified by experts as “positive samples”
- Compare each retrieved product to the positive samples!
    - 1) Mean Euclidean distance from the “positive” labelled samples
    - 2) Min. Euclidean distance from the “positive” labelled samples
    - 3) Max. Euclidean distance from the “positive” labelled samples

- Directly predict the rating which the user may give to the product!!!!

Product Recommendation

- Based on cumulative ratings by all customers
- Based on similarity with product that user is currently viewing
  - based on customer statistics
  - based on product features - K-nearest neighbors
- Based on user's personal history
  - which products is the user likely to like???

Product Ratings based on Features



How much rating will a particular user give this camera out of 5?

Probably depends on features!

Which features does the user like?

Source: Flipkart website

Feature Selection

- The user has exactly 5 options: 1, 2, 3, 4 or 5 stars!
- Her choice depends on the different features of the product!
- But she may consider some features to be more important than others !
- Which features determine her vote?

Feature Selection

- The user has exactly 5 options: 1, 2, 3, 4 or 5 stars!
- Her choice depends on the different features of the product!
- But she may consider some features to be more important than others !
- Which features determine her vote?

Company	Color	Resolution	Video Rate	Price	Her Rating
C1	Black	10 MP	25 fps	$200	2
C1	White	15 MP	25 fps	$250	2
C2	White	12 MP	30 fps	$250	4
C1	Black	15 MP	30 fps	$300	3
C2	Black	20 MP	25 fps	$400	3
C2	White	12 MP	50 fps	$500	5
C2	Black	15 MP	30 fps	$250	????

Feature Selection

- The user has 5 exactly options: 1, 2, 3, 4 or 5 stars!
- Her choice depends on the different features of the product!
- But she may consider some features to be more important than others !
- Which features determine her vote?

Company	Color	Resolution	Video Rate	Price	Her Rating
C1	Black	10 MP	25 fps	$200	2
C1	White	15 MP	25 fps	$250	2
C2	White	12 MP	30 fps	$250	4
C1	Black	15 MP	30 fps	$300	3
C2	Black	20 MP	25 fps	$400	3
C2	White	12 MP	50 fps	$500	5
C2	Black	15 MP	30 fps	$350	4

Feature Selection

- The user has 5 exactly options: 1, 2, 3, 4 or 5 stars!
- Her choice depends on the different features of the product!
- But she may consider some features to be more important than others !
- Which features determine her vote?

Company	Color	Resolution	Video Rate	Price	Her Rating
C1	Black	10 MP	25 fps	$200	2
C1	White	15 MP	25 fps	$250	2
C2	White	12 MP	30 fps	$250	4
C1	Black	15 MP	30 fps	$300	3
C2	Black	20 MP	25 fps	$400	3
C2	White	12 MP	50 fps	$500	5
C2	Black	15 MP	30 fps	$350	4

Decision Tree for Feature Selection

- Which features does she consider as important while rating?
- Let's look at her history of rating 100 cameras!

Rating	Count
1	21
2	24
3	18
4	20
5	17

Overall, Count=100

Rating	Count
1	15
2	18
3	10
4	5
5	6

Company = C1, Count=54

Rating	Count
1	6
2	6
3	8
4	15
5	11

Company = C2, Count=46

Rating	Count
1	15
2	20
3	13
4	12
5	10

Color=Black, Count=70

Rating	Count
1	6
2	4
3	5
4	8
5	7

Color=White, Count=30

Decision Tree for Feature Selection

- Which features does she consider as important while rating?
- Let's look at her history of rating 100 cameras!

Rating	Count
1	21
2	24
3	18
4	20
5	17

Overall, Count=100

Rating	Count
1	15
2	18
3	10
4	5
5	6

Company = C1, Count=54

Rating	Count
1	6
2	6
3	8
4	15
5	11

Company = C2, Count=46

Rating	Count
1	15
2	20
3	13
4	12
5	10

Color=Black, Count=70

Rating	Count
1	6
2	4
3	5
4	8
5	7

Color=White, Count=30

Which feature is more important for ratings - company or color???

What's a discriminative feature?

- Company ={C1, C2}, Price = real number, Y = {LOW (1-3), HIGH (4-5)}

	COMPANY=C1	COMPANY=C2	
#(Y=LOW)	43	20	63
#(Y=HIGH)	11	26	37
Total	54	46	100

What's a discriminative feature?

- Company ={C1, C2}, Price = real number, Y = {LOW (1-3), HIGH (4-5)}

	Price<300	**Price >=300**	
#(Y=LOW)	45	18	63
#(Y=HIGH)	25	12	37
Total	70	30	100

What's a discriminative feature?

- Company ={C1, C2}, Price = real number, Y = {LOW (1-3), HIGH (4-5)}

	Price<500	Price >=500	
#(Y=LOW)	55	8	63
#(Y=HIGH)	35	2	37
Total	90	10	100

What's a discriminative feature?

- Prob(Y = HIGH | COMPANY = C1) = 11/54 ~ 0.2 [Easy to decide]
- Prob(Y = HIGH | COMPANY = C2) = 26/46 ~ 0.55

- Prob(Y = HIGH | PRICE < 300) = 25/70 ~ 0.36
- Prob(Y = HIGH | PRICE >= 300) = 12/30 = 0.4

- Prob(Y = HIGH | PRICE < 500) = 35/90 ~ 0.4
- Prob(Y = HIGH | PRICE >= 500) = 2/10 ~ 0.2 [Easy to decide][Very few examples]

What's a discriminative feature?

- Prob(Y = HIGH | COMPANY = C1) = 11/54 ~ 0.2 [Easy to decide]
- Prob(Y = HIGH | COMPANY = C1) = 26/46 ~ 0.55

COMPANY: good feature

- Prob(Y = HIGH | PRICE < 300) = 25/70 ~ 0.36
- Prob(Y = HIGH | PRICE >= 300) = 12/30 = 0.4

PRICE<300: bad feature

- Prob(Y = HIGH | PRICE < 500) = 35/90 ~ 0.4
- Prob(Y = HIGH | PRICE >= 500) = 2/10 ~ 0.2 [Easy to decide][Very few examples]

PRICE<500: doubtful feature