IR UNIT – 4 (Probabilistic Retrieval and Language Modeling & Related Methods, Categorization & Filtering) – END-SEM PYQ Answers

- **NOV/DEC 2022**

Q3) a) Explain Categorization and Filtering with any two detailed Examples. [9]

Categorization and Filtering are two important tasks in Information Retrieval and Machine Learning.
1. Categorization: Categorization (or *classification*) is the task of assigning a document to one or more predefined categories based on its content.

Process:

1. Preprocessing (tokenizing, stemming, stop-word removal)
2. Feature extraction (bag-of-words, TF-IDF, word embeddings)
3. Apply classifier (Naive Bayes, SVM, Neural Networks)
4. Predict class label

Example: Email Spam Classification

- Goal: Categorize emails into Spam and Not Spam.
- Steps:
    - Extract features like frequency of suspicious words (“free”, “win”, “offer”).
    - Train model using labeled emails.
    - New emails classified based on learned patterns.

Use case: Gmail, Yahoo Mail spam detection.

2. Filtering: Filtering is the task of allowing only *relevant* documents to pass through according to user preferences or predefined criteria.

Types:

- Rule-Based Filtering (keyword-based)
- Content-Based Filtering (user profile based)
- Collaborative Filtering (similar users' choices)

Example: News Recommendation System

- Users prefer topics like “Technology” or “Sports".
- System filters news articles by matching:
    - Keywords
    - Reading history
    - User profiles
- Only relevant news items are shown to each user.

Use case: Google News, Flipboard.

Example (Alternative): Web Content Filtering

- Organizations filter out websites with malware, adult content, or restricted topics.
- Based on rules, URL categories, or machine learning models.

Conclusion: Categorization assigns documents to predefined classes, whereas filtering selects only the documents relevant to a user or rule. Both are essential for managing large document collections.

b) Explain the Information-Theoretic Model in detail. [8]

The Information-Theoretic Model in IR and machine learning is based on the idea that information can be quantified in terms of entropy, uncertainty, and probability distributions.

1. Entropy

Entropy measures the uncertainty or randomness in a dataset.

$$H(X) = -\sum p(x) \log_2 p(x)$$

Higher entropy → more uncertainty.

2. Mutual Information (MI)

MI measures how much information one variable gives about another.

$$I(X; Y) = H(X) - H(X \mid Y)$$

Used for:

- Feature selection
- Understanding term–document relationships

3. KL Divergence

Measures the difference between two probability distributions.

$$D_{KL}(P \mid\mid Q)$$

Used in:

- Language modeling
- Ranking
- Evaluating model fit

4. Application in IR

- Term weighting using information gain
- Query-document similarity using divergence methods

- Ranking using probability distributions (e.g., LM, BM25 variants)
- Feature selection using MI or IG

5. Interpretation

- A term provides more "information" if it reduces uncertainty.
- Rare but meaningful terms contribute more to ranking.
- Models aim to minimize uncertainty in prediction and maximize shared information.

Conclusion: The Information-Theoretic Model provides mathematical tools like entropy, MI, and KL divergence to measure information, relevance, and uncertainty in IR and machine learning systems.

Q4) a) Explain Probabilistic Classifiers & Generalized Linear Models. [9]

1. Probabilistic Classifiers: Probabilistic classifiers assign a document to the class with the highest posterior probability.

$$P(class \mid document)$$

They output probability values, not just labels.

Common Probabilistic Classifiers:

i) ***Naive Bayes Classifier***

Based on Bayes' theorem:

$$P(C \mid X) = \frac{P(X \mid C)P(C)}{P(X)}$$

Assumes features are independent.
Used widely for text classification (spam detection, sentiment analysis).

ii) ***Logistic Regression (Probabilistic Form)***

Outputs probability using the logistic function:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-W^T X}}$$

Used for binary classification.

Advantages

- Simple
- Fast
- Works well for high-dimensional data (text)

2. Generalized Linear Models (GLMs)

GLMs extend linear regression to handle:

- Non-normal distributions
- Non-linear relationships

Components of GLMs:

1. **Linear Predictor:**

$$\eta = w^T x$$

2. **Link Function:**
   Converts linear predictor to mean of distribution.
   Examples:
   - Logit → Logistic Regression
   - Probit → Probit Model
   - Identity → Linear Regression
3. **Probability Distribution:**
   Chosen from exponential family (Gaussian, Bernoulli, Poisson).

Examples of GLMs

- Logistic Regression (Bernoulli)
- Poisson Regression (Count data)
- Softmax Regression (Multiclass classification)

Importance

- Provide a unified statistical framework
- Support classification and regression
- Widely used in IR, NLP, and predictive modeling

b) Describe Language Models and Smoothing. [8]

1. Language Models (LMs)

A Language Model estimates the probability of a word sequence.

$$P(w_1, w_2, \ldots, w_n)$$

Used in:

- Speech recognition
- Machine translation
- Query likelihood in IR

a) Unigram Model

Assumes word independence: $P(w_1, w_2) = P(w_1)P(w_2)$

b) Bigram / Trigram Models

Use previous 1 or 2 words for prediction: $P(w_n \mid w_{n-1})$

c) LM in IR: Query Likelihood Model

Rank documents by: $P(query \mid document \quad LM)$

Document with highest probability is ranked first.

2. Smoothing in Language Models

Smoothing adjusts probability estimates to avoid zero probabilities for unseen words.

i) Laplace (Add-One) Smoothing

Adds 1 to each term frequency: $P(w \mid d) = \frac{tf(w,d)+1}{|d|+V}$

Simple but may over-smooth.

ii) Add-k Smoothing

Adds a fractional constant k: $tf + k$

iii) Jelinek-Mercer Smoothing (Linear Interpolation)

Combines document and collection model: $P(w \mid d) = \lambda P(w \mid d) + (1 - \lambda) P(w \mid C)$

iv) Dirichlet Prior Smoothing

Most effective method: $P(w \mid d) = \frac{tf(w,d) + \mu P(w \mid C)}{|d| + \mu}$

Where μ is prior strength.

Purpose of Smoothing

- Avoids zero probabilities
- Stabilizes LM performance
- Helps short documents rank properly
- Improves retrieval accuracy

► **MAY/JUNE 2023**

Q3) a) Explain Categorization and Filtering with any two detailed Examples. [9] → DONE

b) Explain the Information-Theoretic Model in detail. [8] → DONE

Q4) a) Explain probabilistic Classifiers & Generalized Linear Models. [9] → DONE

b) Describe Language Models and Smoothing. [8] → DONE

► **NOV/DEC 2023**

Q3) a) Explain Categorization and Filtering with any two detailed Examples of each. [8] → DONE

b) Write a short note on Generalized Linear Models. [9] → DONE

Q4) a) Explain Probabilistic Classifiers, Information-Theoretic Model in detail. [9] → DONE

b) Describe Language Models and Smoothing. [8] → DONE

► **MAY/JUNE 2024**

Q3) a) Explain Probabilistic Classifiers & Generalized Linear Models. [9]

1. Probabilistic Classifiers

Probabilistic classifiers assign each document or instance to the class that has the highest posterior probability given the feature values.

$$P(Class \mid Document)$$

They output not only a class label but also the confidence probability.

Key Ideas

- Based on Bayes theorem
- Use probability distributions
- Useful in text classification, spam detection, recommendation systems

---

i) ***Naive Bayes Classifier***

The most common probabilistic classifier.

$$P(C \mid X) = \frac{P(X \mid C)P(C)}{P(X)}$$

Assumption: Features are conditionally independent.
Steps:

1. Compute prior $P(C)$
2. Compute likelihood $P(x_i \mid C)$
3. Choose class with maximum posterior $P(C \mid X)$

Used in: Spam filtering, Sentiment analysis, Document categorization.

ii) ***Logistic Regression***

A probabilistic binary classifier using sigmoid function.

$$P(y=1 \mid x)=\frac{1}{1+e^{-w^T x}}$$

Outputs the probability that instance belongs to class 1.

Advantages of Probabilistic Classifiers

- Simple and fast
- Can handle high-dimensional data
- Provide uncertainty/confidence scores

2. Generalized Linear Models (GLMs)

GLMs extend linear regression to allow:

- Non-normal response variables
- Non-linear relationships
- A link function connecting linear predictor to mean

Components of GLMs

1. Linear predictor: $\eta = w^T x$
2. Link function:
    - Logit → Logistic Regression
    - Probit → Probit Model
    - Log → Poisson Regression
3. Distribution:
   Chosen from exponential family (Bernoulli, Poisson, Gaussian)

Examples

- Logistic Regression → Classification
- Poisson Regression → Count prediction
- Softmax Regression → Multiclass classification

GLMs provide a unified mathematical framework used across IR, ML, NLP, and statistics.

b) Explain the Information-Theoretic Model in detail. [8]

The Information-Theoretic Model applies concepts from information theory—like entropy and mutual information—to understand and measure relevance in IR and ML.

1. Entropy: Entropy measures uncertainty or randomness in data.

$$H(X)=-\sum p(x)\log\ p(x)$$

Higher entropy → more randomness.
Lower entropy → more predictability.

2. Mutual Information (MI): MI measures how much information one variable provides about another.

$$I(X;Y) = H(X) - H(X \mid Y)$$

Used for:

- Term selection
- Feature selection
- Understanding term–document relevance

3. KL Divergence: KL divergence measures how one probability distribution differs from another.

$$D_{KL}(P \mid\mid Q)$$

Used in:

- Language models
- Ranking
- Model comparison

Lower divergence → distributions are similar.

4. Application in IR

- Terms that reduce document uncertainty are more informative
- Ranking based on probabilistic similarity
- Feature weighting via information gain
- Query-document matching via divergence (LM-based retrieval)

5. Interpretation

- Rare but meaningful terms carry high information
- Frequent generic terms carry less information
- The model maximizes shared information between query and document

Q4) a) Explain Categorization and Filtering with any two detailed Examples. [9]

Categorization and filtering are fundamental IR tasks for organizing and delivering relevant information.

1. Categorization: Categorization assigns a document to one or more predefined classes.

Process

1. Document preprocessing (tokenization, stopwords, stemming)
2. Feature extraction (TF-IDF, bag of words)
3. Apply classifier (Naive Bayes, SVM, Neural networks)
4. Assign class label

Example: Email Spam Classification

- Emails labeled as Spam or Ham
- Features include suspicious words (“win”, “free”, “offer”)
- Classifier learns from historical labeled emails
- New emails categorized automatically

Used by Gmail, Outlook.

2. Filtering: Filtering selects only the documents that meet user preferences or rules.

Types

- Rule-based
- Content-based
- Collaborative filtering

Example: News Recommendation System

- User prefers topics: “Sports”, “Tech”
- System filters incoming news articles
- Matches keywords, user history, and content preferences
- Only relevant articles shown

Used by Google News, Inshorts.

Example: Web Content Filtering

- Blocks adult/malicious sites
- Uses URL categories and ML classification

b) Describe Ranking with Language Model. [8]

Language Model (LM)–based ranking is a retrieval approach where documents are ranked based on the probability of generating the query.

1. Query Likelihood Model

For each document, build a document language model.

$$P(Q \mid D)$$

Rank documents by descending probability.

2. Unsmoothed Estimation

$$P(w \mid D) = \frac{tf(w, D)}{|D|}$$

Fails when a query term does not appear in D → probability becomes 0.

Hence smoothing is required.

3. Smoothing Methods

*a) **Jelinek–Mercer Smoothing***

Linear interpolation:

$$P(w \mid D) = \lambda P_{ml}(w \mid D) + (1 - \lambda) P(w \mid C)$$

Mixes document model with collection model.

*b) **Dirichlet Prior Smoothing***

$$P(w \mid D) = \frac{tf(w, D) + \mu P(w \mid C)}{|D| + \mu}$$

Most commonly used; very stable.

*c) **Add-One and Add-k Smoothing***

Adds constant to every term frequency.

4. Ranking Steps

For each document:

1. Estimate smoothed LM
2. Compute probability of generating query
3. Multiply probabilities of all query terms
4. Rank documents by score

5. Advantages

- Strong theoretical foundation
- Handles term rarity well
- Performs better than classical TF-IDF in many tasks

► **NOV/DEC 2024**

Q3) a) Explain categorization and filtering with any two detailed examples. [7]

Categorization and Filtering are two major tasks in Information Retrieval used to organize and deliver relevant information.

1. Categorization (Classification): Categorization is the process of assigning documents to predefined categories based on their content.
It uses machine learning models such as Naive Bayes, SVM, Neural Networks, etc.

Steps in Categorization

1. Preprocessing (tokenization, stemming)
2. Feature extraction (TF-IDF, BoW)
3. Apply classifier
4. Predict category

Example: Email Spam Classification

- Emails are categorized into Spam or Not Spam.
- Features: suspicious words like *free*, *offer*, *win*.
- Trained classifier identifies patterns in spam messages.
- New emails are automatically categorized.

Used in: Gmail, Outlook.

2. Filtering: Filtering delivers information that matches user needs or predefined rules.
It removes irrelevant content and shows only items of interest.

Types

- Rule-based filtering
- Content-based filtering
- Collaborative filtering (user similarity)

Example: News Recommendation System

- User is interested in *Technology* and *Sports* news.
- System examines article keywords, reading history, and user profile.
- Only relevant articles are shown to the user.

Used in: Google News, Flipboard.

Conclusion: Categorization assigns documents to classes, while filtering selects only useful information. Both help manage large-scale information efficiently.

b) Describe passage retrieval and ranking with example. [5]

Passage Retrieval retrieves relevant sections or passages of a document rather than the whole document.
Useful when long documents contain only small relevant parts.

Steps in Passage Retrieval

1. Segment documents into passages (paragraphs or fixed-length windows).
2. Index passages instead of whole documents.
3. Match query against each passage.
4. Score and rank passages by relevance.
5. Return top-ranked passages or map them back to the parent document.

Ranking of Passages

Common ranking approaches:

- TF-IDF cosine similarity
- BM25 score
- Language Model probabilities

Example: Query: *"symptoms of diabetes"*

Document contains:

- Passage 1: About causes
- Passage 2: About treatment
- Passage 3: Symptoms
    - Frequent urination
    - Excessive thirst

Passage 3 receives highest ranking score and is retrieved.

Benefits

- More accurate than whole-document retrieval
- Useful in QA systems, medical IR, legal IR

c) Explain the Information-Theoretic Model in detail. [5]

The Information-Theoretic Model uses entropy, information gain, and probability distributions to measure relevance in IR.

1. Entropy

Measures the uncertainty in a random variable.

$$H(X) = -\sum p(x) \log\ p(x)$$

Lower entropy → more predictable data.

2. Mutual Information (MI)

Measures how much information one variable gives about another.

$$I(X;Y) = H(X) - H(X \mid Y)$$

Used for:

- Feature selection
- Term weighting

3. KL Divergence

Measures difference between two probability distributions.

$$D_{KL}(P \mid\mid Q)$$

In IR: measures similarity between document model and query model.

4. Application in IR

- Ranking documents
- Selecting informative terms
- Comparing query-language models and document models
- Improving relevance scoring

Conclusion: The information-theoretic model provides a mathematical and probabilistic foundation for measuring information, relevance, and uncertainty in IR systems.

Q4) a) Explain probabilistic Classifiers & Generalized Linear Models. [7]

1. Probabilistic Classifiers

Probabilistic classifiers predict the class with the maximum posterior probability.

$$P(Class \mid Document)$$

i) ***Naive Bayes Classifier***

Uses Bayes theorem with independence assumption.

$$P(C \mid X) = \frac{P(X \mid C)P(C)}{P(X)}$$

Widely used in:

- Spam classification
- Sentiment analysis

ii) ***Logistic Regression***

A discriminative probabilistic model.

$$P(y=1 \mid x)=\frac{1}{1+e^{-w^T x}}$$

Outputs class probabilities instead of just labels.

2. Generalized Linear Models (GLMs)

GLMs extend linear regression by using:

1. Linear predictor: $\eta = w^T x$
2. Link function (logit, probit, log)
3. Probability distribution (Bernoulli, Poisson, Gaussian)

Examples

- Logistic regression (binary classification)
- Poisson regression (count prediction)
- Softmax regression (multiclass)

GLMs provide a unified statistical method for classification in IR and ML.

b) Explain Relevance Feedback Technique with suitable diagram. [5]

Relevance Feedback improves retrieval by learning from user feedback on retrieved documents.

Steps

1. User submits query
2. System retrieves initial documents
3. User marks some as *relevant* or *non-relevant*
4. System modifies the query vector
5. Retrieves improved results

Rocchio Algorithm (Vector Model)

$$Q_{new} = \alpha Q + \beta \frac{1}{|D_r|} \sum D_r - \gamma \frac{1}{|D_{nr}|} \sum D_{nr}$$

Where:

- D_r = relevant docs
- D_{nr} = non-relevant docs

Diagram

```
User Query → Initial Retrieval → User Feedback
     ↑                                ↓
Improved Query  ← Updated Results ← System Learns Relevance
```

Advantages

- Better ranking
- Personalized results

Adaptation to user intention

c) Describe Language Models and Smoothing. [5]

1. Language Models (LMs)

A LM assigns a probability to a sequence of words. $P(w_1, w_2 \ldots w_n)$

In IR (Query Likelihood Model): $Rank(D) = P(Q \mid D)$

2. Types of Language Models

- Unigram LM: assumes word independence
- Bigram/Trigram LM: uses previous words

3. Smoothing

Smoothing avoids zero probabilities for unseen words in a document.

a) Laplace/Add-One Smoothing

Adds 1 to term counts: $P(w \mid D) = \frac{tf(w, D) + 1}{\mid D \mid + V}$

b) Jelinek–Mercer Smoothing

Linear interpolation: $P(w \mid D) = \lambda P_{ml}(w \mid D) + (1 - \lambda) P(w \mid C)$

c) Dirichlet Prior Smoothing

Most effective: $P(w \mid D) = \frac{tf(w, D) + \mu P(w \mid C)}{\mid D \mid + \mu}$

Importance

- Handles rare terms
- Stabilizes probability estimation
- Improves retrieval accuracy

NOTE: Please verify all answers before referring.