Week 7 Reading Notes: NLP 1 - Distributions & Key Research Directions

Name: Ananya and Apoorva

Overview

Week 7 covers four main topics:

1. Choosing appropriate regression analysis methods
2. Statistical distributions and Poisson Random Fields
3. Climate change data analysis
4. Introduction to NLP with Count Vectors and TF-IDF

---

1. Choosing the Right Regression Analysis

Source: Statistics By Jim

Types of Regression Models

Linear Regression

- For continuous dependent variables with linear relationships
- Assumptions: linearity, homoscedasticity, independence, normality
- Example: Predicting house prices

Logistic Regression

- For binary outcomes (yes/no, 0/1)
- Output: Probability between 0 and 1
- Example: Customer churn prediction, spam detection

Polynomial Regression

- For non-linear curved relationships
- Uses polynomial terms (x^2, x^3)
- Risk of overfitting with high degrees

Poisson Regression

- For count data (non-negative integers)
- Assumption: Variance equals mean
- Example: Number of customer visits, accident counts

Negative Binomial Regression

- For count data with overdispersion
- Use when variance exceeds mean

Decision Framework

Questions to ask when choosing regression:

1. What type is my dependent variable?
2. What is the relationship shape?
3. Are model assumptions met?
4. Is there overdispersion in count data?

---

2. Distributions & Poisson Random Fields

Key Distributions

Normal Distribution

- Bell curve, symmetric
- Parameters: mean and standard deviation
- Common in natural phenomena

Poisson Distribution

- Models count of events in fixed interval
- Parameter: λ (average rate)
- Mean equals variance
- Good for rare events

Binomial Distribution

- Number of successes in n trials
- Parameters: n (trials), p (probability)

Exponential Distribution

- Time between events in Poisson process
- Example: Time until next customer arrival

Poisson Random Fields

Concept:

- Extension of Poisson processes to spatial domains
- Models discrete events distributed across space or time
- Used for feature selection in high-dimensional data

Dynamic Feature Models:

- Features can appear or disappear over time
- Allows sparse, dynamic feature sets
- Applications: time-series, gene expression, topic modeling

Key Innovation:

- Traditional models assume fixed features
- Poisson Random Fields enable flexible feature selection
- Better for high-dimensional sparse data

---

3. Climate Change Data Analysis

Types of Climate Data

- Temperature records
- Precipitation measurements
- Sea level data
- Ice core samples
- Satellite observations
- Atmospheric CO_2 levels

Statistical Challenges

Temporal Autocorrelation

- Climate measurements correlated over time
- Requires time-series models (ARIMA, SARIMA)

Spatial Autocorrelation

- Nearby locations have similar climates
- Requires spatial statistics

Non-stationarity

- Climate patterns change over time
- Requires detrending or differencing

Missing Data

- Historical records incomplete
- Requires imputation techniques

Regression Applications

- Linear regression for temperature trends
- Polynomial regression for non-linear patterns
- Poisson regression for extreme event counts
- Machine learning for climate forecasting

Key Points

- Climate change detection requires 30+ years of data
- Natural variability must be separated from human impact
- Multiple data sources strengthen conclusions

---

4. Introduction to NLP: Count Vectors and TF-IDF

Text Preprocessing Pipeline

Step 1: Text Cleaning

- Remove punctuation and special characters
- Convert to lowercase
- Handle contractions

Step 2: Tokenization

- Split text into individual words
- Example: "I love NLP" → ["I", "love", "NLP"]

Step 3: Stop Word Removal

- Remove common words: "the", "is", "at", "a"
- Reduces dimensionality

Step 4: Stemming/Lemmatization

- Stemming: Crude chopping (running → run)
- Lemmatization: Dictionary-based (better → good)

Count Vectors (Bag of Words)

Concept:

- Represent text as vector of word frequencies
- Each unique word is one dimension
- Value is count of word in document

Example:

Document 1: "I love dogs"
Document 2: "I love cats"

Vocabulary: [I, love, dogs, cats]
Count Vectors:
Doc1: [1, 1, 1, 0]
Doc2: [1, 1, 0, 1]

Advantages:

- Simple and interpretable
- Captures word importance by frequency

Disadvantages:

- Ignores word order
- Treats all words equally
- High dimensionality

TF-IDF (Term Frequency - Inverse Document Frequency)

Concept:

- Weighs words by importance across documents
- Down-weights common words, up-weights rare words

Formula:

TF-IDF = TF × IDF

TF = (Count of word in document) / (Total words in document)
IDF = log(Total documents / Documents containing word)

Intuition:

- High TF: Word appears frequently in this document
- High IDF: Word is rare across all documents
- High TF-IDF: Word is important to this specific document

Advantages:

- Reduces impact of common words
- Highlights discriminative terms
- Better for classification and search

Disadvantages:

- Still ignores word order
- More complex than count vectors
- Requires entire corpus for calculation

Applications

- Text classification (spam detection, sentiment analysis)
- Information retrieval (search engines)
- Document similarity (plagiarism detection)

Python Implementation

Count Vectorizer:

```
from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer()
X = vectorizer.fit_transform(corpus)
```

TF-IDF Vectorizer:

```
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(corpus)
```

Connections Between Topics

Regression + Climate Data:

- Poisson regression for extreme weather events
- Linear regression for temperature trends

Distributions + NLP:

- Word frequencies follow power-law distributions
- Topic modeling uses Dirichlet distributions

NLP + Regression:

- TF-IDF features as predictors
- Sentiment scores predicting outcomes

Dynamic Features + NLP:

- Poisson Random Fields for topic evolution
- Sparse feature selection in text data

Key Takeaways

Regression Analysis

- Choose regression based on dependent variable type
- Always validate model assumptions
- Start simple, add complexity as needed

Statistical Distributions

- Poisson for count data with rare events
- Poisson Random Fields enable dynamic models
- Distribution choice affects model performance

Climate Data

- Long-term trends require careful analysis
- Account for temporal and spatial autocorrelation
- Multiple data sources strengthen conclusions

NLP Basics

- Count Vectors: Simple frequency-based representation
- TF-IDF: Weights words by importance

- Preprocessing is critical for success

Tools & Libraries

- scikit-learn: CountVectorizer, TfidfVectorizer, regression models
- NLTK: Natural language toolkit
- spaCy: Industrial NLP
- statsmodels: Advanced statistical models
- pandas: Data manipulation