11912007, yahya jabary

MovieLens 100K Dataset by GroupLens Research Project at the University of Minnesota. Data was collected through the MovieLens web site (movielens.umn.edu) during the seven-month period from September 19th, 1997 through April 22nd, 1998.

Data is preprocessed: users who had less than 20 ratings or did not have complete demographic information were dropped.

- `u.user`: demographic user information
- `u.item`: movie information
- `u.data`: ratings
- `u.info`: total counts: "943 users, 1682 items, 100000 ratings"
- `u.genre`: ordinal encoding of genres: "unknown|0, Action|1, Adventure|2, … Western|18"
- `u.occupation`: list of user occupations: "administrator, artist, doctor, educator, engineer, … writer"
- `allbut.pl`, `mku.sh`: utility scripts to generate training and test sets and unzip the tar
- `u[1-5|a].base`, `u[1-5|a].test`: 80%/20% training and test sets created by `u1.base` and `u1.test` (5-fold cross-validation), also with alternative splits `ua.` and `ub.`

schema

- `u.user`:
  - 0: `user_id` - (can be joined on ratings table, starting from 1)
  - 1: `age`
  - 2: `gender` - binary m/f
  - 3: `occupation` - (lookup in `u.occupation`)
  - 4: `zip_code`
- `u.item`:
  - 0: `movie_id` - (can be joined on ratings table, starting from 1)
  - 1: `movie_title`
  - 2: `release_date` - date DD-MMM-YYYY
  - 3: `video_release_date` - date DD-MMM-YYYY
  - 4: `IMDb_URL`
  - 5-23: `genre` - binary encoded, multiple genres can apply
- `u.data`:
  - 0: `user_id`
  - 1: `movie_id`
  - 2: `rating` - 1-5
  - 3: `timestamp` - unix seconds since 1/1/1970 UTC

---

analysis

univariate analysis

- `u.user`: user
  - 0: `user_id`
    - 943 unique values
  - 1: `age`
    - min: 7, max: 73
    - left skewed towards younger ages
    - mean: 32.97, median: 30.00, std dev: 11.56, skewness: 0.73, kurtosis: -0.17
    - quantiles: 25th: 24.00, 75th: 40.00
  - 2: `gender`
    - 74% male, 26% female
  - 3: `occupation`
    - 21 unique values
    - 22% student, 11% other, 9% educator, 8% engineer, …, 1% lawyer, 0.9% none, 0.9% salesman, 0.5% doctor, 0.3% homemaker
  - 4: `zip_code`
    - 795 unique values
- `u.item`: movie
  - 0: `movie_id`
    - 1682 unique values
  - 1: `movie_title`
    - 1664 unique values
  - 2: `release_date`
    - 9 missing values
    - ranges from 1922 to 1998
    - 26% on january 1st
  - 3: `video_release_date`
    - 100% missing values
  - 4: `IMDb_URL`
    - 13 missing values
  - 5-23: `genre`
    - 40% drama, 30% comedy, 26% action, 22% thriller, 20% romance, 14% adventure, …, 1.4% fantasy, 0.8% documentary
- `u.data`: rating
  - 0: `user_id`
  - 1: `movie_id`
  - 2: `rating`
    - each user has rated at least 20 movies
    - ranges 1-5
    - right skewed towards higher ratings
    - mean: 3.53, median: 4.00, std dev: 1.13, skewness: -0.51, kurtosis: -0.41
    - quantiles: 25th: 3.00, 75th: 4.00
  - 3: `timestamp`
    - ranges 1997-09-20 to 1998-04-22

bivariate analysis

- correlation encoded: 0.877 `IMDb_URLxmovie_title`, 0.555 `ChildrensxAnimation`, 0.451 `AdventurexAction`, 0.417 `MusicalxAnimation`, 0.381 `MusicalxChildrens`, 0.323 `ActionxSci-Fi`, …, 0.054 `agexrating`
- `genderxoccupation`: male (22% student, 10.8% engineer, 9.9% programmer, 9.4% other, …), female (22% student, 14% other, 11% librarian, 10% administrator, …)
- `genderxrating`: male (3.53 mean, 4.00 median, 1.11 std), female (3.53 mean, 4.00 median, 1.17 std) → chi-squared statistic of 0.53 and p-value of 0.9705
- `agexrating`: rating increases with age except for 20-30 y.o.: <20 y.o. (3.55 mean, 1.15 std), 20-30 y.o. (3.44 mean, 1.16 std), 30-40 y.o. (3.57 mean, 1.10 std), 40-50 y.o. (3.57 mean, 1.10 std), 50+ y.o. (3.68 mean, 1.02 std) → pearson correlation coefficient of 0.0545
- `occupationxrating`: highest by mean: none (3.78), lawyer (3.74), doctor (3.69), educator (3.67), …, homemaker (3.30), healthcare (2.90)

---

ethics

"The user may not use […this…] forany commercial or revenue-bearing purposes without first obtaining permission from a faculty member of the GroupLens Research Project at the University of Minnesota."

- privacy: user re-identification with demographics data / osint (open-source intelligence) → need for anonymization, salted hashes, homomorphic encryption
- gdpr: they would be required to give explicit consent under GDPR for personal data collection and processing
- bias: sample bias and age of data may lead to bad recommendations
- harm: recommender systems are "minimal risk" systems under the EU AI Act, and aren't regulated
- harm: collection of <18 y.o. data requires parental consent and need special protection
- harm: if used for high risk models, data is biased towards a specific age, sex, occupation group and isn't inclusive

---

hypotheses

hypothesis: rating prediction (regression task)

- hypothesis: "matrix factorization algorithms outperform memory-based collaborative filtering for predicting user ratings"
- independent variable: algorithm type (SVD vs User-KNN)
- dependent variable: rating prediction accuracy
- control conditions: same training/test data split, same hyperparameter optimization procedure, same computing environment
- performance indicator: RMSE (root mean square error) and MAE (mean absolute error)
- scale type: ratio scale (ratings 1-5)
- experimental design: random 80/20 train/test split, stratified by user, 5-fold cross-validation for robust results, grid search for hyperparameter optimization on validation set, statistical significance testing of results

hypothesis: cold-start problem (classification task)

- hypothesis: "content-based features provide better recommendations for new users than demographic features alone"
- independent variable: feature set (content vs demographic)
- dependent variable: recommendation quality
- control conditions: same set of new users, same recommendation algorithm, same number of features
- performance indicator: hit rate@K, NDCG@K
- scale type: ordinal (ranking quality)
- experimental design: hold out subset of users as "new users", train models using different feature sets, generate top-K recommendations, compare ranking metrics

non-testable hypotheses

- "users are satisfied with recommendations" (requires explicit feedback)
- "other user demographics beyond age/gender influence rating" (limited demographic data, no session data)
- "ui interactions influence rating behavior" (no implicit feedback data)

experimental considerations

- i. data quality: handle missing values, remove outliers, check for data imbalance, normalize features
- ii. validation strategy: use stratified sampling (sampling from each cluster) for classification, time-based splitting for recommendation tasks, k-fold cross-validation where appropriate
- iii. statistical testing: paired t-tests for comparing models, bootstrap sampling for confidence intervals, effect size calculations, confidence intervals, significance testing, reproducibility
- iv. real-world simulation: time-based validation, cold-start scenarios, sparse data conditions, considering compute and memory constraints, scalability of algorithms, robustness to noisy data

---

theory: data science

data science

- = data driven insights
- interdisciplinary field: cs, stats, domain
- challenges: getting data, overcoming assumptions, communication, managing client expectations

data science process

- 1 – ask research question
  - define variables, metrics, build hypothesis
- 2 – get the data
  - sample, preprocess, ensure privacy

- 3 – explore the data
  - plot, find patterns and anomalies
- 4 – model the data
  - fit a model, validate
  - bias (inaccurate) vs. variance (overfitting)
- 5 – communicate findings
  - report, visualize
  - correlation $\neq$ causation

crisp-dm

- = cross-industry standard process for data mining
- 1 – business understanding
- 2 – data understanding
- 3 – data preparation
- 4 – modeling
- 5 – evaluation
- 6 – deployment

legal & privacy

- gdpr general data protection regulation (may 2018)
  - personal, sensitive, possible re-identification
- eu data strategy (2020)
  - common european data spaces: sharing data pools within the eu
- eu data governance act (sep 2023)
- eu ai act (may 2024)
  - unacceptable risk class = illegal
  - high risk class = products, vehicles, critical infra, health, safety, law $\rightarrow$ must be audited, monitored
  - transparency risk class = risk of deception $\rightarrow$ must be transparent
  - minimal risk class = common stuff like recommender systems, spam filters
- eu data act (sep 2025)

ethics

- algorithmic bias: discrimination, unfairness
- ethical assessment needs to be tracable (of cause and effect)
- concern levels:
  - epistemic:
    * inconclusive evidence - no certainty using stats
    * inscutable evidence - no correlation between data and conclusion
    * misguided evidance - conclusions are only as reliable as the data
  - normative: (based on values, observer dependent)
    * unfair outcomes of decided actions
    * transformative effects in our social perception

---

theory: experiments

needed in empirical science

experiment

- testable hypothesis = we can explain dependent variable with independent variable(s)
  - assumes cause and effect
  - test against a control group (a/b testing)
- metrics:
  - validity = accuracy of instruments
  - reliability = sanity and consistency of outcome, on multiple repetitions

experiment types

- pilot experiment = checking instruments
- natural experiments = pure observations
- field experiments = experiment env hard to control and replicate (most social sciences)
- controlled experiments = lab experiments, outcome is a depedent variable
- factorial experiments = exhaustive dependent variable search

variables

- dependent variables = outcome
- independent variables = arguments
  - changed per experiment
  - values assigned to it are called "control"
- extraneaous/nuisance/interfering variables = noise
  - set to either be constant or all possible values
- confounder variables = influence both dependent and independent variables
  - ie. gender influences both drug and recovery
- latent variables = not directly measurable, hidden from observation

statistical testing

- test null hypotheses by translating them into statistics
- statistics = observations of random variables from known distributions
  - univariate analysis = explains single variable.
  - bivariate analysis = explains relationships / how changes in one variable relate to changes in another.
- statistical inference = making a conclusion about unseen population from a sample
- hypothesis testing:
  - similar to "proof by contradiction" - prove that the probability of a proposition is very low
  - does not prove H_0 but limits the likelihood of H_1 being false based on some level of significance α or p-value
- sampling distribution:
  - estimate distributions analytically
  - use non/parametric statistics to estimate params
  - mean of samples approaches normal distribution as size increases, irrespective of population distribution (central limit theorem)
- tests:
  - z-test: compare difference in means
  - t-test: sampling distribution of diference of means

---

theory: machine learning

data types

- qualitative / categorical data:
  - nominal = unique names
  - ordinal = also have order (sorting)
- quantitative / numerical data:
  - interval = also can be measured and compared on a scale (addition, subtraction)
  - ratio = also have an absolute zero point (multiplication, division)

normalization

- categorical:
  - 1-hot-encoding = map to 0 array with one flag bit
  - distance encoding = map ordinal-data to integers
- numerical:
  - min-max = map to 0;1
    * $z_i = (x_i - \min(X))/(\max(X) - \min(X))$
  - z-score = map distribution to mean 0, std dev 1
    * $z_i = \frac{x_i - \mu}{\sigma}$
  - binning = map value ranges to discrete numbers

missing values

- a) deletion: remove attribute or row (only in train-set)
- b) imputation: don't leak data when reconstructing
  - categorical: `NA` as label, regression, clustering, knn
  - numerical: mean, median, regression, clustering, knn
- not allowed to influence results

sampling

- randomize data first
- stratification = make sure each class is equally represented in all sets
- data-leakage = data from train-set influencing test-set
- validation-set = subset of train-set to tune hyperparameters
- holdout = 80/20 train and test split
- k-fold cross val = split data in k same-sized parts, 1 part for test-set, remaining parts for train-set, repeat k times
- leave-one-out cross val = $k = n$
- leave-p-out cross val = choose unique subset of size p, use this subset for test-set, remaining data for train-set, repeat (too expensive!)
- bootstrapping = sample with replacement, use final sample (bootstrap set) for test-set, remaining data (out-of-bag set) for train-set, repeat

contingency table

- predicted positive + actual positive = true positive
- predicted positive + actual negative = false positive (error I)
- predicted negative + actual positive = false negative (error II)
- predicted negative + actual negative = true negative

confusion matrix

- table of predicted vs. actual for all classes

metrics for classification

- accuracy
  - $\frac{TP+TN}{TP+FP+TN+FN}$
  - correctness of both positives and negatives
- precision
  - $\frac{TP}{TP+FP}$
  - correctness of positives
- specificity
  - $\frac{TN}{TN+FN}$
  - correctness of negatives
- recall, sensitivity
  - $\frac{TP}{TP+FN}$
  - completeness
- balanced accuracy
  - $\frac{\frac{TP}{TP+FN} + \frac{TN}{TN+FP}}{2}$
  - average of precision and specificity
- f1 score
  - $2 \cdot \frac{\text{Prec} \cdot \text{Rec}}{\text{Prec} + \text{Rec}}$

statistical significance testing

- null hypothesis = difference between two systems isn't by chance (like variations in data or randomness in algorithm)
- level of significance α = probability of false negative – ie. 5% p-level means there is a 5% chance that the result is just by chance
- false positives in significance testing likelier if sample is too small