Introduction to Statistical Machine Learning
CSC/DSCC 265/465

Lecture 14: Midterm Review

Cantay Caliskan

Notes and updates

SICSS Summer School (May 9 – May 20)

- **Summer Institute in Computational Social Sciences *(May 9 – May 20)*** at the University of Rochester

- Opportunity to learn, to network, and to listen to exciting speakers, but most importantly: to finish a CSS project in a team (and there is no tuition)

- Topics: ML, image as data, NLP, data visualization, network analysis with applications in CSS

- Advanced UGs interested in PhD, Master’s students, PhD students, postdoctoral students (from different fields) are welcome!

- **Check the link for more information and application (Deadline: March 18): https://sicss.io/2022/rochester/**

SICSS

UNIVERSITY of ROCHESTER

SICSS Summer School (May 9 – May 20)

Midterm

- Practice midterm posted!
- **Midterm will be in-class on March 16, Wednesday. Please arrive 10 minutes early!**
- **<u>Eight (8)</u>** questions in total (**80 points** in total = each question is worth **10 points**)
- **<u>Distribution of questions:</u>**
  - ***~3 Math*** questions (matrix algebra + probability)
  - ***~5 Theory*** questions (ML)
    - <u>~One (1)</u> of these questions might be an exploratory question about a relevant topic we haven't studied in class
- Review will cover the important concepts, but it will not provide an exhaustive coverage

Revisiting: Topics we covered

- Vectors and matrices
- Matrix algebra
- Probability
- Statistical concepts (distributions, definitions on matrices)
- ML basics (cross-validation, bias/variance, regularization, overfitting/underfitting)
- Basics for supervised learning (algorithms, cost functions, optimization)
- Gradient descent
- Maximum likelihood estimation
- Logistic regression (two-class and multi-class)
- Practical examples on NLP
- Model performance (accuracy and goodness-of-fit)
- Regularized linear models
- Clustering (kmeans(++), DBSCAN, OPTICS, GMM – soft/hard clustering, optimization)
- Dimensionality reduction (PCA and SVD)

Matrix Algebra

Concepts: Matrix algebra

- **Matrices were useful in storing data, optimization, calculating similarities, distances…**
- **<u>Vectors and matrices</u>**
  - **Difference between *scalar*, a *vector*, and a *matrix***
    - ***Scalar***: ***1 x 1*** matrix
    - ***Vector***: ***m x 1*** matrix
    - ***Matrix***: ***m x n*** matrix
  - **Vectors**
    - ***Magnitude***: Usually calculated by Euclidean distance
    - ***Direction***: Given by the locations in vector space

  - **More about vectors:**
    - Live in an ***n-dimensional*** space (=***n*** is the number of features)
    - We usually deal with feature vectors (=values associated with an observation/unit of analysis)
    - We also analyzed documents in NLP

Concepts: Matrix algebra

- <u>**Vectors and matrices**</u>
  - **Matrix algebra (addition, multiplication, transpose, inverse)**
    - ***Addition***: Add each element (pairwise)
      - Two matrices need to have the same dimensions.
    - ***Multiplication***: Multiply and add the elements by rows (of the ***1st*** matrix) and columns (of the ***2nd*** matrix)
      - Same as dot product [=usually meant for vectors]
      - Multiplication by identity matrix (***I***) -> Provides the same result
      - Multiplying by inverse -> Provides the identity matrix (***I***)
    - ***Transpose***: Rows become columns, columns become rows
      - Helpful for calculating the <u>Variance-Covariance matrix</u>
      - <u>Var-Cov</u> is helpful in various ways including finding optimal coefficients or doing dimensionality reduction

Concepts: Matrix algebra

- **<u>Vectors and matrices</u>**
  - **Matrix algebra (addition, multiplication, transpose, inverse)**
    - ***Inverse***: Helpful in optimization
      - Hard to compute (check ***pseudo-inverse***)
      - Sometimes not computable
        - Sparse matrices
        - Matrices with no determinant
      - Multiplying with inverse gives the identity matrix -> $\boldsymbol{AA^{-1} = I}$
  - **Special matrices**
    - Symmetric
    - Skew-symmetric
    - Identity
    - Diagonal

Probability

Concepts: Probability

- **Probability concepts:**
  - ***Sample space, events***
    - What are the possible outcomes? All probabilities of independent events sum to 1.
    - Conditional events do not sum to 1.
  - ***Independent events***
  - ***Marginal probability, joint probability, conditional probability***
    - Joint: Two events happening at the same time
    - Conditional: Probability given that something has already happened
  - ***Probability density function***
    - Discrete or continuous (example: balls vs. temperature)
- **Schools of statistics**
  - ***Bayesian vs. Frequentist statistics***
  - Difference: ***Power of sample (Frequentist)*** vs. ***Power of sample+priors (Bayesian)***

Concepts: Probability

- **<u>Probability concepts:</u>**
  - ***Bayesian Statistics***
    - **<u>Bayes’ (or Bayesian) Theorem</u>**
      - Relationship between the conditional probabilities of two events
    - **MLE vs MAP**
      - Maximum Likelihood Estimation vs. Maximum A Posteriori

$$P(A \mid B) = \frac{P(B \mid A) P(A)}{P(B)}$$

More definitions

- ***[Probability] Expectation/Mean***
- ***Matrix decomposition***
- ***Matrix definiteness***
- ***Gaussian Distribution***
  - Bivariate
  - Multivariate

ML Basics

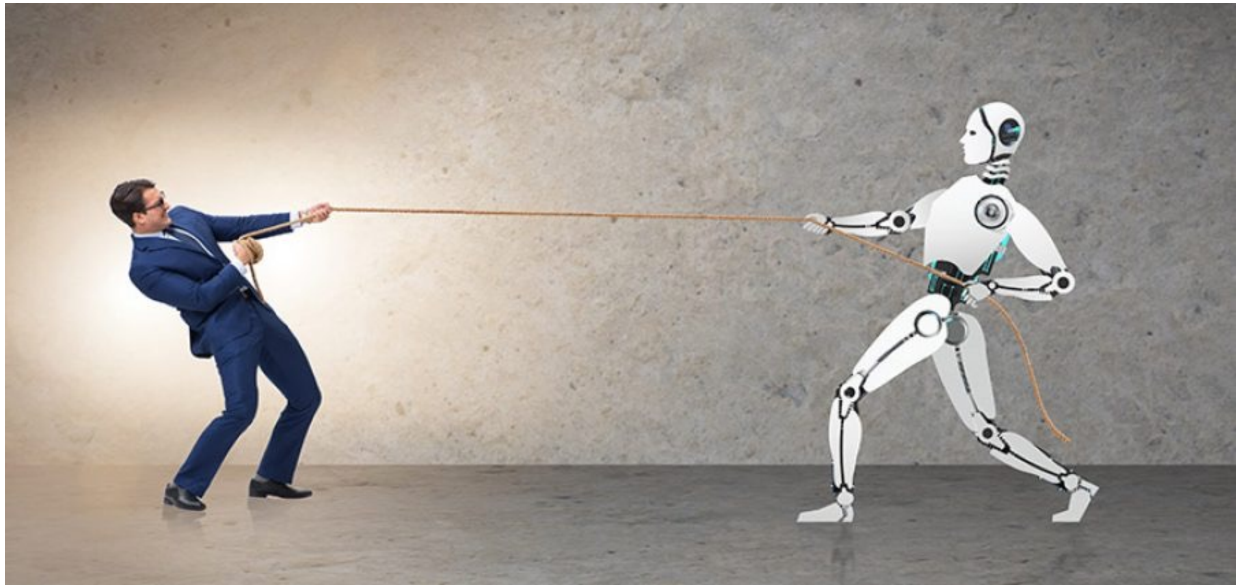
ML Basics

- ***Cross-Validation***
  - A technique used to select the appropriate algorithm (not the coefficients
  - Usually used when you want to test the fragility of the algorithm and/or you have little data
  - Train/test CV ; K-Fold CV ; Leave-one-out CV (LOOCV)
- ***Bias and Variance***
  - Do you want bias or variance in your model?
  - Complex models -> Less bias, more variance
  - Basic models -> more bias, less variance
- ***Underfitting and Overfitting***
  - ***Overfitting*** on the training dataset -> **<u>Variance</u>** in the test dataset
  - ***Underfitting*** on the training dataset -> **<u>Bias</u>** in the test dataset
  - ***Best model*** -> When training and test dataset have comparable errors
- ***Regularization*** (we will come to that later…)

ML Basics

- ***Goodness-of-Fit***
  - <u>Idea</u>: Quantifying the performance of a model
    - R2
    - Accuracy
    - Specificity
    - Sensitivity
    - F1-Score
    - MSE
    - Root-MSE
    - MAE
    - MAPE
    - Confusion Matrices

Supervised Learning



Supervised Learning

- ***Need to have:***
  - Feature vectors + ***label***
  - Or: Independent variables + ***dependent variable***
- Idea: You can teach the computer a pattern in a guided/supervised way
- Goal: Explanatory or predictive analysis
- ***Models:***
  - Categorical data OR continuous data
  - Linear regression, logistic regression
  - Regularized linear regression
- ***Cost Functions:*** ← **Goal**: Minimizing the cost
  - Convex (= closed-form solutions possible)
    - Example: Linear Regression -> SSE
  - Non-Convex (= iterative optimization is necessary) ← **Gradient descent** was one of them
    - Example: Logistic Regression -> Cross-Entropy

Gradient Descent

Gradient Descent

- ***Gradient Descent***
  - An iterative optimization algorithm that can be applied to any supervised learning problem with a defined cost function
  - Intuition: How does the cost function look like?
- ***Gradient***
  - Slope of the tangent at a particular point on the cost function curve
- ***Learning rate***
  - How fast you are willing to move along the curve
- ***Gradient problems***
  - Vanishing gradient
  - Exploding gradient
- ***Types***
  - Stochastic GD, Batch GD, Mini Batch GD

Maximum Likelihood

Maximum Likelihood

- ***Parametric vs. Non-Parametric Models***
  - Are the paremeters given or calculated?
- Idea: Can we estimate the parameters of a distribution to fit the model in a 'better' way?

- ***Estimating MLE***
  - Maximizing the log-likelihood

- ***MLEists***
  - Bayesians with flat priors

Classification

Classification: Logistic Regression

- ***Two-class vs. Multi-class***
  - Binomial logistic regression vs. multinomial logistic regression
  - There are other algorithms, as well…
- ***Generative vs. Discriminative Algorithms***
  - Features of the observation vs. border between observations
  - Generative (=probabilistic), discriminative (=probabilistic/non-probabilistic)
- ***Activation Function***
  - Sigmoid function
  - Comparison with linear regression
- ***Optimization***
  - Iterative learning logistic regression: Gradient Descent
  - Cost Function: Cross-Entropy
- ***Log-Odds vs. Probability***
  - Results need to be converted into probability

Classification: Empirical Example

- ***Natural Language Processing (NLP)***
  - Sentiment analysis
  - Tokenization
  - POS-Tagging
  - Stemming
  - Lemmatization
  - Count-Vectorization
  - TF-IDF-Vectorization

- Idea: Convert the text into a numerical form so that you can make predictions

UNIVERSITY of ROCHESTER

Regularization

Regularization

- ***Solution to:***
  - Overfitting/underfitting
  - Bias/Variance
- **Idea: Make the model generalizable**
  - Regularization vs. Generalization dilemma
- ***Philosophically connected to:***
  - Occam's Razor
  - Parsimonious explanations
- ***Math:***
  - L2, L1, L1+L2, L^p regularization
  - Applications in linear models (Ridge, Lasso, ElasticNet)
- ***Question:***
  - Should we discard some of the features? (=model selection/identification strategy)
  - Should we decrease the importance of some of the features? (=shrinkage)

Unsupervised Learning

Unsupervised Learning

- **Idea:**
  1) Finding patterns in data
  2) Dimensionality reduction
- ***Simple Goal:*** Find clusters. Minimize intra-cluster distances. Maximize inter-cluster distances.

- ***Different clustering methods*** have emerged, because of:
  - Convergence issues
  - Distribution/Shape of Data
  - Different density
  - Cluster size

- ***Clustering families:***
  - Centroid-based (i), density-based (ii), distribution-based (iii), hierarchical (iv)

Unsupervised Learning

- ***Kmeans(++):***
  - Centroid-based clustering
  - Assumption: Gaussian distribution in data
  - Random initialization (-> comes with problems)
  - No convergence-guarantee to global minimum (=each initialization leads to a different result)
  - Faster
- ***DBSCAN:***
  - Density-based clustering
  - Assumption: Clustering can be inferred from local density
  - Constant density parameter (-> comes with problems)
  - Discards some of the data
  - Different densities lead to problems
  - Slower

Unsupervised Learning

- ***OPTICS:***
  - Density-based clustering
  - <u>Assumption</u>: Clustering can be inferred from local density
  - Dynamic density parameter (-> fixes an issue with DBSCAN)
  - Slower
- ***GMM:***
  - Distribution-based clustering
  - <u>Assumption</u>: MLE can be used to identify different distributions
  - Probabilistic approach
  - Expectation-Maximization (EM)
  - Slower

- ***Other issues:*** Choosing the optimal ***k*** (theory vs. Math dilemma), hard vs. soft clustering

Dimensionality Reduction

Dimensionality Reduction

- ***Idea:***
  - Transform a high-dimensional space to a low-dimensional space
  - ***Curse of dimensionality***
  - Feature elimination vs. feature extraction
- ***Methods:***
  - PCA and SVD
- ***PCA:***
  - Transform the data into a lower-dimensional space
  - Maximize the variance of the projected data
  - Connection to Var-Cov matrix, eigenvalues, eigenvectors
  - Assumes linear relationship between variables
  - Creates independent principal components (=helpful)

Do you have any questions?