UVA CS 4774: Machine Learning

S6: Lecture 28: Review Students' Week QA

Dr. Yanjun Qi

University of Virginia
Department of Computer Science

Linear Regression

Understanding the Objective (Loss) Function

- **MSE vs. MAE vs. SSE:** Why do we prefer squaring errors over absolute values, and what is the physical meaning of the 1/2 coefficient?
- **Matrix vs. Summation:** Clarifying the derivation and transition between summation notation and vector/matrix notation.
- **Interpretation:** Differences between "Loss," "Cost," and "Objective" functions.

Generalization & Data Strategy

- **Train/Test Splits:** How to determine the optimal split ratio (e.g., 80/20) and why training performance implies nothing about generalization.
- **Diagnostics:** Identifying overfitting vs. underfitting through residual patterns or loss discrepancies.
- **Bias:** Handling distribution shifts between training and testing data.

Feature Engineering & Model Assumptions

- **Multicollinearity:** How correlated features (e.g., house size vs. bedrooms) affect coefficients and interpretation.
- **Suitability:** Pre-modeling checks to confirm if a dataset is actually linear and how to handle outliers.

Gradient Descent & SGD for Linear Regression

Algorithm Selection: Closed-Form vs. Iterative

- **The Decision Matrix:** When to use Normal Equation vs. GD vs. SGD vs. Mini-batch based on dataset size (*N*) and feature count (*D*).
- **Trade-offs:** Computational complexity of matrix inversion (and "Full Rank" issues) vs. iterative convergence speed.

Hyperparameter Tuning (The "How-To")

- **Learning Rate (α):** Strategies for initialization, dynamic adjustment, and diagnosing "bad" rates (overshooting vs. slow convergence).
- **Batch Size:** Determining the optimal batch size for stability vs. speed.
- **Initialization:** How starting points impact the final result.

Convergence & Topography

- **Convexity:** Defining convex functions, saddle points, and why they guarantee (or don't guarantee) Global Minima.
- **Escaping Minima:** Can GD/SGD get stuck in local minima, and how do we mitigate this?

Linear Prediction with Regularization

Basis Function Selection & Tuning

- How do we choose between polynomial, RBF, or Splines, and how are hyperparameters (e.g., RBF centers/widths) determined?
- Can we mix different types of basis functions, and how does degree/complexity relate to overfitting?
- What are the specific trade-offs regarding model flexibility and computational cost for these functions?

Regularization Mechanics (L1 vs. L2)

- What are the practical and geometric differences between Ridge (L2) and Lasso (L1), specifically regarding feature sparsity?
- When should we use Elastic Net over Ridge or Lasso, and how do we tune the balance?
- How does regularization strictly relate to "shrinking" coefficients and matrix invertibility?

Hyperparameters & Bias-Variance

- How does the regularization parameter (λ) specifically impact the bias-variance trade-off?
- Does increasing regularization always help generalization, or can it increase error in specific cases?

KNN and Theory

KNN Implementation & Scalability

- How do we choose the optimal *k* to balance high variance (small *k*) and high bias (large *k*)?
- How are computational costs/latency handled in production given KNN's lazy learning nature?
- How does the "curse of dimensionality" and feature scaling affect distance metrics?

Validation Methodologies

- Why/when do we prefer *k*-fold Cross-Validation over a single train/val split, and how do we choose *k* for the folds?
- Is data "wasted" in validation splits, and does the model see all data eventually?
- Is it possible for Validation Score to be higher than Training Score, and what does that imply?

Model Selection & Generalization

- How do we interpret validation curves to identify the "middle ground" between underfitting and overfitting?
- What is the fundamental difference between parametric (Linear Reg) and non-parametric (KNN) approaches regarding data usage?
- What strict criteria define unreliable vs. inaccurate estimation?

Bias Variance Tradeoff

Diagnosing Model Performance

- How do we practically identify high bias vs. high variance using learning curves (training vs. validation error)?
- How do we locate the "sweet spot" or ideal ratio between bias and variance when we don't know the true distribution?
- What are the concrete strategies to fix underfitting vs. overfitting (e.g., regularization vs. more data)?

Regularization Nuances (Lasso, Ridge, Elastic Net)

- When should we prioritize Lasso (L1) over Ridge (L2), and how does L1 geometrically yield sparse models/zero coefficients?
- What are the computational cost differences between methods, and is standardization of features strictly necessary?
- Is it always safer to use Elastic Net, and how do we handle correlated variables (grouping effect)?

K-Nearest Neighbors (KNN) & Complexity

- How does the choice of *k* specifically impact the bias-variance tradeoff and overfitting?
- How do we handle distance metrics (e.g., weighted, Hamming) and tie-breaking in classification?
- What is the computational complexity of KNN during testing (e.g., sorting costs)?

Probability Review & Maximum Likelihood Estimation

The Logic of Log-Likelihood

- Why do we maximize the **log**-likelihood ($\ln L$) rather than the raw likelihood?
- Is MLE simply analogous to minimizing error, and what makes it "consistent/efficient"?

Connecting MLE to Linear Regression

- Why is minimizing Squared Error (OLS) equivalent to maximizing likelihood under Gaussian noise assumptions?
- How does the cost function change if we assume a non-Gaussian distribution (e.g., Laplace)?

Distributions & Estimations

- How does MLE naturally extend from discrete (Coin flips/Bernoulli) to continuous (Gaussian) distributions?
- Why is the sample proportion ($\frac{x}{n}$) considered the best estimate for parameters in simple scenarios?
- What if the MLE parameter isn't "ideal" for the specific data—are there other metrics?

Logistic and NN

The Bias-Variance Tradeoff & Model Complexity

- **Quantifying Error:** How do we quantify $noise^2$ versus definable errors when calculating the Expected Prediction Error (EPE)?
- **Complexity:** Why does increasing model complexity lead to a "U-shaped" test error curve, and how do we determine the "best" model from this graph?
- **Diagnostics:** In practical workflows, how do we distinguish between high bias (underfitting) and high variance (overfitting) to decide the next step?

Maximum Likelihood Estimation (MLE) Intuition

- **Probability vs. Loss:** Why does "maximizing likelihood" correspond to finding the best parameters, and how does this mathematically relate to minimizing Squared Error in regression?
- **The Log Transformation:** Why is the log-likelihood ($\ln L(\theta)$) universally preferred over raw likelihood? Is it purely for computational ease (sums vs. products) or convexity?
- **Assumptions:** What happens when MLE assumptions (like independence or correct model specification) are violated?

Logistic Regression Mechanics

- **Link Functions:** How does the log-odds (logit) transformation bridge linear regression and probabilities bounded between $[0,1]$?
- **Optimization:** When fitting logistic models, what are the trade-offs between Gradient Descent and Newton's Method (Hessian computation)?
- **Decision Boundaries:** How does the model determine the specific cutoff point (e.g., 0.5) for classification?

NN and Deep Learning

Architecture & Representation

- **Universal Approximation:** What does it mean for a NN to have the "Universal Approximation Property," and how does adding hidden layers specifically enable the modeling of complex, non-linear functions?
- **Activation Choices:** What criteria guide the choice between Sigmoid, Tanh, and ReLU (e.g., diminishing returns, non-convexity), and why is Softmax preferred for multi-class tasks?
- **Design:** How do we practically decide the number of hidden layers and neurons per layer?

Training Dynamics (Backpropagation)

- **The Chain Rule:** How exactly do "local gradients" (like $\frac{\partial E}{\partial z}$) combine layer-by-layer to update earlier weights?
- **Non-Differentiability:** How does backpropagation handle points where activation functions are non-differentiable (e.g., ReLU at 0)?
- **Initialization:** If we skip Xavier initialization and use random numbers, how does this specifically affect training convergence?

Optimization & Loss Functions

- **Loss Selection:** Why is Cross-Entropy mathematically better suited for classification than Mean Squared Error (MSE), and how does it relate to the Sigmoid function?
- **Gradient Strategy:** How does Stochastic Gradient Descent (SGD) differ from Batch/Mini-batch in terms of noise and convergence speed?
- **Generalization:** What is the specific difference between overfitting and misfitting in the context of deep networks?

CNN

Architecture Components: Convolution & Pooling

- How do Convolutional layers differ from Fully Connected layers, and why are they better for images?
- What is the specific utility of Max Pooling/Subsampling versus Strides?
- How does the hierarchical structure lead to abstract features?

Mechanisms: Invariance & Weight Sharing

- How does weight sharing create translation invariance/equivariance?
- Why do we reuse the same filter, and how does that handle object movement?
- What is the difference between translation invariance and subsampling?

Implementation: Hyperparameters & Training

- How do we determine the optimal number of layers, filter depths, and kernel sizes?
- Can CNNs be applied to non-image data (e.g., video) or optimized for large images?
- How does Backpropagation function specifically within CNNs?

PCA, Feature Selection

Feature Selection vs. Feature Extraction

- What is the fundamental difference in how they transform or choose features?
- What are the main trade-offs, and how do we decide which is better for a dataset?

PCA Mechanics & Limitations

- Does PCA fail on data with nonlinear feature interactions (linearity assumption)?
- What does the First Principal Component actually represent, and why is centering mandatory?
- How does PCA handle small datasets?

Comparisons & Advanced Reduction

- How does PCA compare to LDA, t-SNE, or Autoencoders?
- How do we ensure reduction preserves semantic meaning rather than just variance?
- Can we combine multiple reduction methods in one model?

Recent deep learning on Text

Evolution of Architectures: RNNs to Transformers

- Why have we shifted from Recurrent Neural Networks (RNNs) to Transformers/Attention mechanisms?
- How do Transformers handle long-range dependencies and parallelization better than RNNs?
- How will these models continue to evolve in the future?

Transformer Mechanics: BERT vs. GPT

- What distinguishes "Understanding" (Encoder/BERT) tasks from "Generation" (Decoder/GPT) tasks?
- How do pre-training objectives (e.g., Masked LM) improve performance compared to previous methods?
- Why are positional encodings necessary in the absence of recurrence?

Text Representation & Embeddings

- Why do Embeddings (Word2Vec) and Contextual Embeddings outperform Bag-of-Words?
- When should we use CBOW vs. SkipGram, and what does "closeness" in embedding space really imply?
- How do we handle Out-of-Vocabulary (OOV) words or misspellings?

Generative Classification

Conceptual Foundations

- What is the fundamental difference between Generative models and Discriminative models (e.g., Logistic Regression)?
- What is the relationship between the Bayes Classifier (theoretical best), Generative Bayes, and Naive Bayes?
- How do we practically apply Bayes Rule for classification tasks?

Naive Bayes & Independence

- Why does Naive Bayes perform well even when the "independence" assumption is violated by correlated features?
- How do we handle zero probabilities using smoothing (Laplace/Lidstone)?
- How are probabilities efficiently stored and looked up in modern implementations?

Gaussian Discriminant Analysis (LDA/QDA)

- How do we estimate class conditionals $P(X|Y)$ (e.g., Gaussian assumptions) in practice?
- How do the decision boundaries differ between LDA (shared covariance) and QDA (class-specific covariance)?
- How sensitive are these Gaussian estimates to outliers and scaling?

NaiveBC on Text

The "Naive" Independence Assumption

- Why does the classifier perform well on text even though the independence assumption is clearly violated (context matters)?
- What are the specific risks or failure modes when this assumption is violated in practice?
- How do we handle words that never appear in the training set (Zero Probability/OOV) and how does smoothing help?

Model Variations: Multivariate Bernoulli vs. Multinomial

- When should we prefer one over the other (e.g., document length, vocabulary size)?
- Why does Multinomial generally outperform Bernoulli in text classification tasks?
- Does the "Bag of Words" approach lose too much semantic information (word order/syntax)?

Generative vs. Discriminative Classifiers

- What is the fundamental difference in what they model ($P(X|Y)$ vs $P(Y|X)$)?
- When would we prefer a Generative model (NB) over a Discriminative one?
- Why is NB considered "generative" if we only use it to classify?

SVM

The Kernel Trick & Feature Spaces

- Intuition: How do we map to infinite dimensions without explicit computation, and why doesn't this cause immediate overfitting?
- Selection: How do we choose the right kernel (RBF vs. Polynomial) and hyperparameters (e.g., γ, degree)?
- Preprocessing: Why is feature scaling/normalization crucial before applying kernels?

Margins and Generalization

- Why does maximizing the margin theoretically lead to better generalization on unseen data?
- Hard vs. Soft Margin: How do slack variables (ξ) and the C parameter control the trade-off between margin size and misclassification?
- What is the difference between functional and geometric margins?

Optimization & The Dual Formulation

- Why do we solve the Dual problem (Lagrange multipliers) instead of the Primal, and how does this relate to the kernel trick?
- The role of Support Vectors: Why does the model only depend on a few points, and what happens if non-support vectors are moved?
- What does it mean for a kernel matrix to be Positive Semi-Definite (Mercer's Condition)?

Decision Trees, Bagging, and Random Forests

Splitting Criteria and Information Theory

- What is the intuition behind minimizing entropy/impurity, and what does it actually represent?
- How does the algorithm decide which attribute to split on, and does Information Gain favor attributes with many distinct values?
- Why is a greedy approach used, and does it guarantee an optimal tree?

Single Tree Instability and Overfitting

- Why are individual decision trees considered "unstable" and prone to overfitting?
- Why does Bagging work effectively for high-variance models (DTs) but less so for stable models (SVMs/Logistic Regression)?

Mechanisms of Variance Reduction (Random Forest)

- Why does averaging highly overfit trees result in a robust model (Bias-Variance decomposition)?
- How does random feature selection ($m < d$) decorrelate trees, and how does this differ from standard Bagging?
- How do we determine the optimal number of features (m) or trees to prevent overfitting?

Boosting

Bagging (Parallel) vs. Boosting (Sequential)

- What are the fundamental differences between training in parallel (Bagging) vs. sequentially (Boosting)?
- When should one prefer Boosting over Bagging/Random Forest for a given dataset?
- Does the sequential nature of boosting cause efficiency/runtime issues compared to parallel methods?

Boosting Mechanics and Error Handling

- How does the algorithm identify and reweight "hard" examples or residuals to reduce error?
- What problem is Boosting trying to solve that Bagging does not address (Bias vs. Variance)?
- How do modern implementations like XGBoost handle missing values or prevent overfitting to noise?

Bias-Variance Tradeoff in Boosting

- At what point does adding more weak learners shift from reducing bias to overfitting the data?
- How do learning rates and tree depth affect generalization in boosting compared to bagging?
- Why does Boosting reduce bias while Bagging primarily reduces variance?

Clustering (Partitioning & Hierarchical)

Configuration & Optimization of K-Means

- How do we determine the optimal *K* in practice (e.g., Elbow method reliability) to avoid overfitting?
- Beyond random restarts, are there better seeding strategies (like K-means++) to avoid poor local optima?
- How does initialization affect convergence and the stability of the final clusters?

Hierarchical Clustering Decisions

- How do we choose between Single, Complete, and Average linkage, and what are the trade-offs regarding noise/outliers?
- Is there a principled way to determine where to "cut" the dendrogram to form the final clusters?
- How do time complexity and scalability compare to K-Means for large datasets?

Evaluation & Assumptions

- Since clustering is unsupervised, how do we formally evaluate success (internal vs. external criteria) without ground truth?
- How do we handle datasets where K-Means assumptions (spherical clusters, Euclidean distance) fail?
- ~~What are the indications to prefer Density-based (DBSCAN) or GMM over centroid-based methods?~~

Reinforcement Learning

Core Concepts & Trade-offs

- How does RL fundamentally differ from Supervised Learning regarding loss functions (risk min vs. discounted return)?
- How is the Exploration vs. Exploitation trade-off implemented, and what are the risks of leaning too far one way?
- Why must an agent sometimes choose a "worse" immediate action to secure long-term gains?

~~**Rewards, States, and The Environment**~~

- What are the best practices for "Reward Shaping" to avoid unintended or "hacked" behaviors?
- How do we handle real-world scenarios where the Markov property (MDP) is violated?
- How does the agent/environment interaction complicate learning compared to static datasets?

~~**Algorithms & Learning Methods**~~

- What are the practical differences between Monte Carlo, Dynamic Programming, and Temporal Difference learning?
- When do we transition from tabular methods to Deep RL or approximate methods?
- Why is Q-learning considered "off-policy" and how does that differ from on-policy methods?