Recap

Course logistics

General idea of machine learning

- Data consists of input x and correct output y
- Model predicts $\hat{y} = wx + b$
- Loss function compares prediction with data: e.g. $L(w, b) = \left(y - \hat{y}(w, b)\right)^2$
- Minimize L to get best w, b
  - $w^*, b^* = \arg\min_{w,b} L(w, b)$

Linear algebra review

- Vectors (column by default)
- Linear combinations (affine, convex)
- Span and linear independence
- Inner products and norms
- Linear subspace and basis

Recap

Linear algebra review

- Matrices (inner/outer products, transpose)
- Matrix rank, inverse, interpretation (diagonal and orthogonal)
- Singular value decomposition
- SVD interpretation
- Eigendecomposition (as special case of SVD and in general)
- Relationship between SVD and eigendecomposition
- Norms (vector and matrix) and matrix definiteness

Recap

Math for ML review

- Taylor expansion for functions of multiple variables
- Quadratic forms and their visualization

Recap

Quadratic forms $x^\top A x$

- Visualization depending on definiteness of A
- Non-symmetric A, definiteness of $A^\top A$
- Quadratic forms and their visualization

Convexity

- Definition, visual interpretation
- Special cases: quadratic forms

Optimality conditions

- First and second order conditions

Ordinary least squares

- Problem setup, including multivariate case
- Square loss

Recap

Ordinary least squares

- Optimizing the square loss function
- Painful partial derivatives
- Towards more convenient matrix derivatives

Recap

Linear Regression (Theory)

- Matrix derivatives
- Optimizing the least squares objective
- Pseudoinverse
- Multiple output linear regression

Linear Regression (Coding example)

- Synthetic data generation
- Ordinary least squares
- Polynomial features
- Model selection

Recap

Ordinary Least Squares: Coding Example

- Model selection, validation sets
- Basic generalization concepts
- Cross validation, ridge regression

Probability Review

- Basic terminology
- Discrete and continuous RVs: pdf, pmf, cdf
- Joint, marginal, conditional distributions
- Chain rule, independence, Bayes' rule
- Expected value, moments
- Entropy, KL divergence, mutual information
- Vector notation
- Gaussian distribution

Recap

Gaussian distributions

- Geometric interpretation in the multivariate case
- Transformations of Gaussian random variables
- Block notation, marginalization

Recap

Gaussian distributions

- Block notation, marginal and conditional distributions

Probabilistic models

- Likelihood, log likelihood
- Maximum likelihood estimation: interpretation of OLS
- Prior and posterior distribution over parameters
- Maximum a posteriori estimation: interpretation of L2 regularization
- Connection between square loss and Gaussian observation noise

Bayesian inference

- Point estimate vs. full posterior distribution

Recap

Bias variance trade-off

- Theoretical analysis of overfitting and underfitting
- Expected error = Bias^2 + Variance + Irreducible Error
- Revisit example from Colab notebook

Optimization

- Nonlinear least squares
- Gradient descent

Recap

Optimization Algorithms

- Gradient descent
- Momentum
- Nesterov’s accelerated gradient (NAG)

- Preconditioners
- Adaptive gradient (AdaGrad)
- Adam
- Newton’s method

Recap

Stochastic Optimization

- Using a random subset of terms in summation to approximate gradient
- Unbiased estimate of gradient
- Hyperparameter tips (your mileage may vary)

Neural Networks

- Learnable features
- Multi-layer perceptions
- Terminology

Recap

Neural Networks

- Multi-layer perceptrons
- Terminology
- Universal function approximation
- Multivariate chain rule
- Backpropagation

Recap

Neural Networks

- Backpropagation
- Goal: $\frac{\partial L}{\partial \vec{w}_{l,j,\cdot}^{\top}}$
- So far:
  - $\frac{\partial L}{\partial \vec{h}_L} = W_L^{\top} \frac{\partial L}{\partial \hat{\vec{y}}}$
  - $\frac{\partial L}{\partial \vec{h}_l} = W_l^{\top} \frac{\partial L}{\partial \vec{z}_{l+1}}$
  - $\frac{\partial L}{\partial \vec{z}_l} = \begin{pmatrix} g'(z_{l,1}) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & g'(z_{l,n_l-1}) & 0 \end{pmatrix} \frac{\partial L}{\partial \vec{h}_l}$

Recap

Neural Networks

- Computing $\frac{\partial L}{\partial \vec{w}_{l,j,\cdot}^{\top}}$
- Visualizing an MLP
- Gradient issues in training MLPs
    - Math related to exploding and vanishing gradients
    - Weight initialization
    - Weight decay (L2 regularization)
    - Layer normalization

Recap

Neural Networks

- Non-convexity
- Residual connections

Recap

Neural Network Architectures Intro

Classification

Support vector machines

- Goal and geometric interpretation
- Terminology: margin, decision boundary, support vectors
- Mathematical formulation and simplification
- Solution to optimization, weak duality

Recap

Weak duality

- Generalized Lagrangian (penalizing constraint violation)
- Primal and dual optimization problems
- Duality gap $p^* - d^*$

The SVM dual problem

- Writing down the generalized Lagrangian and dual problem
- Solving the inner optimization → simplify the dual problem

Recap

Support Vector Machine

- Margin maximization, geometric intuition and math
- Derived and simplified primal problem
- Derived and simplified dual problem
- Used Slater’s condition to show strong duality
- Obtained expressions for optimal solution and interpreted complementary slackness
- Used kernels to make working with features more efficient / tractable
- Soft-margin SVM

Recap

Support Vector Machine

- Soft-margin SVM
- Reformulation into an unconstrained optimization
- General form of loss functions: 0-1 vs. hinge vs. square

Logistic regression

- Cross-entropy loss