Introduction to Machine Learning

Jonathan Shewchuk

1. Introduction. Classification. Training, validation, and testing. Overfitting and underfitting.

2. Linear classifiers. Decision functions and decision boundaries. The centroid method. Perceptrons.

3. Gradient descent, stochastic gradient descent, and the perceptron learning algorithm. Feature space versus weight space. The maximum margin classifier, aka hard-margin support vector machine (SVM).

4. The support vector classifier, aka soft-margin support vector machine (SVM). Features and nonlinear decision boundaries.

5. Machine learning abstractions: application/data, model, optimization problem, optimization algorithm. Common types of optimization problems: unconstrained, linear programs, quadratic programs. The influence of the step size on gradient descent.

6. Decision theory, also known as risk minimization: the Bayes decision rule and the Bayes risk. Generative and discriminative models.

7. Gaussian discriminant analysis, including quadratic discriminant analysis (QDA) and linear discriminant analysis (LDA). Maximum likelihood estimation (MLE) of the parameters of a statistical model. Fitting an isotropic Gaussian distribution to sample points.

8. Eigenvectors, eigenvalues, and the eigendecomposition of a symmetric real matrix. The quadratic form and ellipsoidal isosurfaces as an intuitive way of understanding symmetric matrices. Application to anisotropic multivariate normal distributions. The covariance of random variables.

9. MLE, QDA, and LDA revisited for anisotropic Gaussians.

10. Regression: fitting curves to data. The 3-choice menu of regression function + loss function + cost function. Least-squares linear regression as quadratic minimization. The design matrix, the normal equations, the pseudoinverse, and the hat matrix (projection matrix). Logistic regression; how to compute it with gradient descent or stochastic gradient descent.

11. Newton's method and its application to logistic regression. LDA vs. logistic regression: advantages and disadvantages. ROC curves. Weighted least-squares regression. Least-squares polynomial regression.

12. Statistical justifications for regression. The empirical distribution and empirical risk. How the principle of maximum likelihood motivates the cost functions for least-squares linear regression

and logistic regression. The bias-variance decomposition; its relationship to underfitting and overfitting; its application to least-squares linear regression.

13. Ridge regression: penalized least-squares regression for reduced overfitting. How the principle of maximum a posteriori (MAP) motivates the penalty term (aka Tikhonov regularization). Subset selection. Lasso: penalized least-squares regression for reduced overfitting and subset selection.

14. Decision trees; algorithms for building them. Entropy and information gain.

15. More decision trees: decision tree regression; stopping early; pruning; multivariate splits. Ensemble learning, bagging (bootstrap aggregating), and random forests.

16. More decision trees: decision tree regression; stopping early; pruning; multivariate splits. Ensemble learning, bagging (bootstrap aggregating), and random forests.

17. Neural networks. Gradient descent and the backpropagation algorithm.

18. The vanishing gradient problem. Rectified linear units (ReLUs). Backpropagation with softmax outputs and cross-entropy loss. Neuron biology: axons, dendrites, synapses, action potentials. Differences between traditional computational models and neuronal computational models.

19. Heuristics for faster training. Heuristics for avoiding bad local minima. Heuristics to avoid overfitting. Convolutional neural networks. Neurology of retinal ganglion cells in the eye and simple and complex cells in the V1 visual cortex.

20. Unsupervised learning. Principal components analysis (PCA). Derivations from maximum likelihood estimation, maximizing the variance, and minimizing the sum of squared projection errors. Eigenfaces for face recognition.

21. The singular value decomposition (SVD) and its application to PCA. Clustering: k-means clustering aka Lloyd's algorithm; k-medoids clustering; hierarchical clustering; greedy agglomerative clustering. Dendrograms.

22. The geometry of high-dimensional spaces. Random projection. The pseudoinverse and its relationship to the singular value decomposition.

23. Learning theory. Range spaces (aka set systems) and dichotomies. The shatter function and the Vapnik–Chervonenkis dimension.

24. AdaBoost, a boosting method for ensemble learning. Nearest neighbor classification and its relationship to the Bayes risk.

25. The exhaustive algorithm for k-nearest neighbor queries. Speeding up nearest neighbor queries. Voronoi diagrams and point location. k-d trees. Application of nearest neighbor search to the problem of geolocalization: given a query photograph, determine where in the world it was taken.

26. Materials

September 3, 2025