CS 439 Exam 02 Cheat Sheet

Linear Algebra & SVD/PCA

Vectors & Spaces

$\mathbf{v} \cdot \mathbf{w} = \sum v_i w_i$ $\|\mathbf{v}\| = \sqrt{\sum v_i^2}$ Linear independence: No vector is a linear combo of others.

Eigenvalues

$A\mathbf{v} = \lambda\mathbf{v}$ λ is eigenvalue, $\mathbf{v}$ eigenvector.

SVD

$X = U\Sigma V^T$, $\Sigma = \text{diag(singular values)}$. Rank-$k$: $X_k = \sum_i^k \sigma_i u_i v_i^T$

PCA

Project data onto eigenvectors of covariance matrix. Sort by eigenvalues, pick top-k.

Probability & Bayes

Conditional Probability

$P(A|B) = \frac{P(A,B)}{P(B)}$ Independence: $P(A,B) = P(A)P(B)$

Bayes Theorem

$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$; $P(B)$ by law of total probability.

Marginal

$P(A) = \sum_i P(A, B_i)$

Naive Bayes Classification

Assumptions

Features $X_1, ..., X_n$ conditionally independent given Y.

Classification Rule

$\hat{y} = \arg\max_y \; P(Y)\prod_i P(X_i|Y)$

Steps

Get prior, likelihood, multiply-score, pick largest.
Laplace smoothing: +1 to all counts for zero probabilities.

Linear Regression

Hypothesis

$h_\theta(\mathbf{x}) = \theta_0 + \theta_1 x_1 + ... + \theta_n x_n = \theta^T \mathbf{x}$

MSE

$\frac{1}{n}\sum(h_\theta(\mathbf{x}) - y)^2$ Normal Equation: $\theta = (X^T X)^{-1} X^T y$

Losses

L2: Squared error, L1: Absolute error, Huber (mix).

Gradient

$\theta_j \leftarrow \theta_j - \alpha(\partial L/\partial\theta_j)$

Model Complexity & Regularization

Bias vs. Variance

Bias (underfit): High train/test error. Variance (overfit): Low train, high test error.

Regularization

Loss: MSE $+ \lambda\sum_j \theta_j^2$ (L2, Ridge). Large λ = more shrinkage. L1 (Lasso): $\lambda\sum_j |\theta_j|$, can zero coefficients.

Choosing λ

Use cross-validation (K-fold, hold out sets). Never penalize intercept θ_0.

Logistic Regression

Sigmoid

$g(z) = \frac{1}{1+e^{-z}}$ Binary output as probability of class 1.

Hypothesis

$h_\theta(x) = g(\theta^T x)$; $P(y = 1|x, \theta)$

Boundary

$h_\theta(x) \geq 0.5 \iff \theta^T x \geq 0$.
Boundary: $\theta_0 + \theta_1 x_1 + ... + \theta_n x_n = 0$.

Loss

Cross-entropy (binary): $L = -y\log(h) - (1-y)\log(1-h)$ For all data: $-\frac{1}{n}\sum_i[y^{(i)}\log(h(x^{(i)})) + (1-y^{(i)})\log(1-h(x^{(i)}))]$

Gradient Desc.

$\theta_j \leftarrow \theta_j - \alpha\frac{1}{n}\sum(h_\theta(x) - y)x_j$

Multiclass Classification

OvA

One-vs-All: Train k binary classifiers, $h_\theta^{(i)}(x)$. Predict class with highest $h_\theta^{(i)}(x)$.

Softmax

$P(y = k|x) = \frac{e^{\theta_k^T x}}{\sum_{j=1}^K e^{\theta_j^T x}}$

MLE

Maximize likelihood over all classes.

Evaluation Metrics

Confusion Matrix

TP = True Pos, FP = False Pos, TN = True Neg, FN = False Neg.

Metrics

Precision: $\frac{TP}{TP+FP}$ Recall: $\frac{TP}{TP+FN}$
Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$ F1: $2\frac{Precision \cdot Recall}{Precision+Recall}$ Specificity: $\frac{TN}{TN+FP}$

Tradeoff

High precision $\rightarrow$ less false positives. High recall $\rightarrow$ less false negatives.

K-means Clustering

Algorithm

Initialize k centers. Assign: $c^{(i)} = \arg\min_j ||x^{(i)} - \mu_j||$. Update $\mu_j = \frac{1}{|C_j|}\sum_{i\in C_j} x^{(i)}$. Repeat until stable.

Loss

$J = \sum_i ||x^{(i)} - \mu_{c(i)}||^2$ (distance to center). Loss always decreases or stays same.

Choosing k

Elbow method: plot loss vs k, look for sharp bend.

K-means++

Pick first center randomly, next picks weighted by distance squared.

Complexity

$O(Imnk)$ where I=iterations, m=points, n=features, k=clusters.

Hierarchical Clustering

Agglom/Divisive

Agglomerative: Start with m clusters, merge closest pairs. Divisive: Start with 1, split.

Linkage

Single: min distance; Complete: max; Average: mean of all pairs.

Dendrogram

Tree shows merge history. Cut at height for k clusters.

Complexity

$O(m^3)$ naive, $O(m^2 \log m)$ with priority queue.

Feature Engineering

Select relevant features (filter, wrapper, embedded). Extract: polynomial, interactions, domain ideas. Normalize: $z = \frac{x-\mu}{\sigma}$. One-hot encoding for categoricals.

Advanced Topics

Cross-Validation

K-fold: Rotate which subset is validation.

Over/Underfitting

High variance (overfit): fits noise; fix with regularization, more data, simpler model. High bias (underfit): misses pattern; fix by adding features, more complex model.

Gradient Tricks

Batch: all data. Stochastic: one at a time. Minibatch: subset.

Parametric vs Non-Parametric

Parametric: fixed parameters; Non-parametric: grow with data.

PCA

Standardize X, covariance $C = \frac{1}{n}X^T X$, decompose, project onto top eigenvectors.

MLE

Choose parameters maximizing probability of observed data.

Common Mistakes

Don't penalize intercept in regularization. Check train/test errors for over/underfitting. Never use test set for training or tuning.

Quick Reference

Bayes: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ Linear: $h = \theta^T x$, MSE $= \frac{1}{n}\sum(h-y)^2$
Gradient: $\theta \leftarrow \theta - \alpha\nabla L$ Regularized: MSE $+ \lambda\sum\theta^2$ Sigmoid: $g(z) = \frac{1}{1+e^{-z}}$ Logistic: $h = g(\theta^T x)$, Loss $= -y\log(h) - (1-y)\log(1-h)$
Precision: $\frac{TP}{TP+FP}$, Recall: $\frac{TP}{TP+FN}$ K-means: $J = \sum||x - \mu_c||^2$