Rishabh Indoria

# Machine Learning Foundations

September 25, 2023

## 1 Week 1

1. Supervised Learning: Regression
    - Find model $f$ such that $f(x^i) \approx g^i$
    - Training Data: $(x^1, y^1), (x^2, y^2), ..., (x^n, y^n)$
    - Loss = $\frac{\Sigma(f(x^i)-y^i)^2}{n}$
    - $f(x) = w^T \cdot x + b$
2. Supervised Learning: Classification
    - $y^i \epsilon -1, +1$
    - Loss = $\frac{\Sigma 1(f(x^i) \neq y^i)}{n}$
    - $f(x) = sign(w^T \cdot x + b)$
3. Validation Data: Choosing the right collection of models is done using validation data.
4. Unsupervised Learning: Dimensionality Reduction
    - Data = $x^1, x^2, ..., x^n$
    - Compress, Explain and Group Data.
    - Encoder $f : R^d \rightarrow R^{d'}$ Decoder: $f : R^{d'} \rightarrow R^d$
    - Goal: $g(f(x^i)) \approx x^i$
    - Loss: $\frac{\left\| g(f(x^i)) - x^i \right\|^2}{n}$
5. Unsupervised Learning: Density Estimation
    - Probabilistic Model
    - $P : R^d \rightarrow R_+$ that sums to 1
    - $P(x)$ is large is $x \epsilon$ Data and low otherwise
    - Loss: $\frac{\Sigma - log(P(x^i))}{n}$

## 2 Week 2

1. Continuity & Differentiability
    - $f : R \rightarrow R$ is continuous if $\lim_{x \rightarrow x^*} f(x) = f(x^*)$
    - Differentiable if $\lim_{x \rightarrow x^*} \frac{f(x)-f(x^*)}{x-x^*} = f(x')$ exists.
    - if $f$ is NOT continuous $\implies$ NOT differentiable
2. Linear Approximation
    - If $f$ is differentiable
    - $f(x) \approx f(x^*) + f'(x^*)(x - x^*)$
    - $f(x) \approx L_{x^*}[f](x)$
3. Higher order Approximation
    - $f(x) \approx f(x^*) + f'(x^*)(x - x^*) + \frac{f''(x^*)}{2!}(x - x^*)^2 + ...$
4. Lines
    - Line through point $u$ along vector $v = x, x = u + \alpha v$, where $u, v, x \epsilon R^d$ and $\alpha \epsilon R$
    - Line through points $u$ and $u' = x, x = u + \alpha(u - u')$

5. Hyper Planes
    - Hyper Plane normal to vector $w$ with value $b = x, w^T \cdot x = b$, where $x, w \epsilon R^d$ and $b \epsilon R$
6. Partial Derivatives & Gradients
    - $f : R^d \to R$
    - $\frac{\delta f}{\delta x}(v) = [\frac{\delta f}{\delta x_1}(v), \frac{\delta f}{\delta x_2}(v), ..., \frac{\delta f}{\delta x_d}(v)]$
    - $\Delta f(v) = [\frac{\delta f}{\delta x}]^T$
7. Multivariate Linear Approximation
    - $f(x) \approx f(v) + \Delta f(v)^T (x - v) = L_v[f](x)$
8. Directional Derivative
    - $D_u[f](v) = \frac{\delta f}{\delta x}(v)^T \cdot u$, at point $v$ along $u$
9. Direction of steepest ascent
    - Find $u \epsilon R^d$, $\|u\| = 1$ & maximize $D_u[f](v)$
    - $u = \alpha \cdot \Delta f(v)$

# 3 Week 3

1. Four Fundamental Sub Spaces
    - Column Space $C(A)$
        - span$(u_1, u_2, ..., u_n)$ = Linear Combination of vectors
        - If $Ax = b$ has a solution, then $b \epsilon C(A)$
        - Rank = number of pivot columns = dim$(C(A))$
    - Null Space $N(A)$
        - $x | Ax = 0$
        - If $A$ is invertible then $N(A)$ only contains zero, and $Ax = b$ has a unique solution.
        - Nullity = number of free variables = dim$(N(A))$
        - If $A$ has $n$ columns, then rank + nullity = $n$
        - Can use Gaussian Elimination to solve for $N(A)$
    - Row Space $R(A)$
        - Column Space of $A^T$
        - Column Rank $dim(C(A))$ = Row Rank $dim(R(A))$
        - $R(A) \perp N(A)$
    - Left Null Space $N(A^T)$
        - $C(A) \perp N(A^T)$
2. Orthogonal and Vector Sub Spaces
    - Orthogonal Vectors, $x \perp y$ if $x \cdot y = x^T y = 0$
    - Orthonormal Vectors, $u \perp v$ and $\|u\| = \|v\| = 1$
3. Projections
    - Projection onto a line
        - $p = \hat{x} a$
        - $e = b - p = b - \hat{x} a$
        - $e \perp a \implies \hat{x} = \frac{a^T b}{a^T a}$
        - Projection matrix $P = \frac{a a^T}{a^T a}$
        - $p = Pb$
        - $P$ is symmetric, $P^2 = P$, Rank $P = 1$
    - Projection onto a subspace
        - Projection of $b$ onto $C(A)$, $Ax = b$

  - $p = A\hat{x}, e = b - A\hat{x}$
  - $e \perp$ every vector in $C(A)$ and $N(A^T) \perp C(A) \implies e \epsilon N(A^T)$
  - Projection Matrix $P = A(A^TA)^{-1}A^T$, $p = Pb$

4. Least Squares

   - Suppose we have a vector $b$ which leads to an inconsistent system $Ax \neq b$
   - Next best thing we do is minimize average error, $E^2 = (Ax - b)^2$
   - $\frac{\delta E^2}{\delta x} = 0 \implies (A^TA)x = A^Tb$

# 4 Week 4

1. Linear and Polynomial Regression

   - Minimize Loss $L(\theta) = \frac{\Sigma(x_i^T - y_i)^2}{2}$
   - Use least squares method $(A^TA)\theta = A^TY$
   - Polynomial Regression
     - Transformed Features: $\hat{y}(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_m x^m = \Sigma\theta_j\phi_j(x), \phi_j(x) = x^j$
     - $\hat{y}(x) = \theta^T\phi(x)$, $(A^TA)\theta = A^TY$
     - Then Proceed as Linear Regression
   - Regularized Loss
     - $\bar{L}(\theta) = \frac{(x_i^T\theta - y_i)^2}{2} + \lambda \left\|\theta\right\|^2$, Regularized Term $= \lambda \left\|\theta\right\|^2$
     - $(A^TA + \lambda I)\theta_{reg} = A^TY$
     - Overfitting → Too small $\lambda$
     - Underfitting → Too large $\lambda$

2. Eigenvalues and Eigenvectors

   - Eigenvalue equation $Ax = \lambda x$
   - $\frac{\delta u}{\delta t} = Au$ can be solved with solutions of the form $u(t) = e^{\lambda t}x$ if $Ax = \lambda x$
   - $(A - \lambda I)x = 0$  
     Characteristic polynomial $|A - \lambda I| = 0$  
     Trace of $A = \Sigma\lambda =$ Sum of diagonal elements of $A$  
     $|A| =$ Determinant of $A = \Pi\lambda$

3. Diagonalization of a Matrix

   - A matrix $A$ is diagonalizable if there exists an invertible matrix $S$ such that $S^{-1}AS = \lambda$, $\lambda =$ Diagonal Matrix
   - $S = \begin{bmatrix} x_1 & x_2 & ... & x_n \end{bmatrix}$, $x_1, x_2, ..., x_n =$ eigenvectors
   - $S^{-1}A^kS = \lambda^k$, $k \geq 1$
   - $Q\lambda Q^T = A$  
     $Q = \begin{bmatrix} q_1 & q_2 & ... & q_n \end{bmatrix}$  
     $q_1 = \frac{x_1}{\left\|x_1\right\|}, q_2 = \frac{x_2}{\left\|x_2\right\|}, ..., q_n = \frac{x_n}{\left\|x_n\right\|}$

4. Fibonacci Sequence $F_k \approx \frac{1}{\sqrt{5}}(\frac{1+\sqrt{5}}{2})^k$

# 5 Week 5

1. Complex Matrices

   - $C^n$: Complex counter part of $R^n$
   - inner product $x \cdot y = \bar{x}^Ty$  
     $\bar{x}^Ty \neq \bar{y}^Tx$  
     $\left\|x\right\|^2 = \bar{x}^Tx$
   - $A^*$ =Conjugate Transpose of $A = \bar{A}^T$

2. Hermitian Matrix
   - $A^* = A$, equivalent of symmetric matrices in complex
   - All Eigenvectors are real and orthogonal
3. Unitary Matrix
   - $U^*U = I$
   - $\left\|Ux\right\| = \left\|x\right\|$
   - $U^{-1} = U^*$
   - $|\lambda| = 1$, where $\lambda$ is any eigenvalue
4. Diagonalization of Hermitian Matrices
   - $A$ is unitary diagonalizable if $A = U\lambda U^*$
   - Any $n \times n$ matrix $A$ is similar to an $n \times n$ upper triangular matrix, $A = UTU^*$
   - If $U_1 = \begin{bmatrix} w_1 & w_2 & ... \end{bmatrix}$ is the matrix then take $w_1 = X_1$, first eigenvector then $w_2 = X_2 - \frac{w_1 \cdot X_2}{\left\|w_1\right\|^2} w_1$

# 6 Week 6

1. Singular Value Decomposition
   - Let $A$ be a real symmetric matrix
     Then all eigenvalues of $A$ are real and $A$ is orthogonally diagonalizable
     $A = Q\lambda Q^T$, $Q^TQ = I$
   - Any real $m \times n$ matrix $A$ can be decomposed to SVD form
     $A(m \times n) = Q_1(m \times m)\Sigma(m \times n)Q_2(n \times n)$, $Q_1^TQ_1 = I$, $Q_2^TQ_2 = I$
     $Q_1 \& Q_2$ are orthogonal
   - $\Sigma = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}$, where $D = \begin{bmatrix} \sigma_1 & 0 & 0 & ... & 0 \\ 0 & \sigma_2 & 0 & ... & 0 \\ 0 & 0 & 0 & ... & \sigma_r \end{bmatrix}$
   - $\sigma_i$ are called singular values and $\sigma_i = \sqrt{\lambda_i}$
     where $\lambda_i$ are eigenvalues of $A^TA$ and $x_i areeigenvectors$
   - Let $y_i = \frac{A_i x_i}{\sigma_i}$
   - $Q_1 = \begin{bmatrix} y_1 & y_2 & ... & ym \end{bmatrix}$, where $y_i$ are eigenvectors of $AA^T = Q_1\Sigma\Sigma^TQ_1^T$ and
     $Q_2 = \begin{bmatrix} x_1 & x_2 & ... & xm \end{bmatrix}$, where $x_i$ are eigenvectors of $A^TA = Q_2\Sigma^T\Sigma Q_2^T$
2. Positive Definite
   - A function $f$ that vanishes at $(0, 0)$ and is strictly positive at other points
   - For $f(x, y) = ax^2 + bxy + cy^2$ to be positive definite
     $a, c > 0$ and $ac > b^2$
   - If $ac = b^2$ then $f(x, y)$ is positive semi-definite($a > 0$) or negative semi definite($a < 0$)
   - If $ac < b^2$ then $(0, 0)$ is saddle point
   - $f(x, y) = v^TAv$, where $v = \begin{bmatrix} x \\ y \end{bmatrix}$ and $A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$
     $|A| < 0 \implies$ saddle point, eigenvalues of $A$ are positive if $f(x, y)$ is positive definite
     $|A| = 0 \implies$ semi definite

# 7 Week 7

1. Principal Component Analysis
   - Start with as many features as you can collect, and then find a good subset of features. Project the data onto a lower dimensional subspace such that Reconstruction error is minimized, Variation of projected error is maximized.
   - Actual: $x_i = \sum_{j=1}^{d}(x_i^Tu_j)u_j$, Projected: $\tilde{x} = \sum_{j=1}^{m} z_{ij}u_j + \sum_{j=m+1}^{d} \beta_j u_j$

- Loss function J = $\frac{1}{n}\sum_{i=1}^{n}||x_i - \tilde{x}_i||^2$
  Differentiating and setting to 0 we get $z_{ij} = x_i^T u_j$ and $\beta_j = \bar{x}_j^T u_j$
- So for a given m dimensional subspace spanned by B = $\{u_1, u_2, ..., u_m\}$ the projected data is
  $\tilde{x}_i = \sum_{j=1}^{m}(x_i^T u_j)u_j + \sum_{j=m+1}^{d}(\bar{x}^T u_j)u_j$
  Loss $J^* = \sum_{j=m+1}^{d} u_j^T C u_j$, $C = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$, where $\{u_1, u_2, ..., u_d\}$ are eigenvectors of $C$.
  For maximizing variance the maximizer is eigenvector of $C$ corresponding to max eigenvalue, the max variance is also equal to this eigenvalue.

2. PCA in higher dimension

   - Suppose D = $\{x_1, x_2, ..., x_n\}$ where $x_i \epsilon R$ and $d >> n$, it would be easier to handle a n by n matrix rather than a d by d matrix.
   - $C = \frac{1}{n}\sum(x_i - \bar{x})(x_i - \bar{x})^T = \frac{1}{n}\sum A^T A$ is a d by d matrix.
   - Since rank($C$)$\leq n \implies (d-n)$ eigenvalues are 0, Hence it is enough to find eigenvectors of $C = \frac{1}{n}AA^T$ which is a n by n matrix.

# 8 Week 8

1. Introduction to Optimization

   - Pillars of ML: Linear Algebra, Probability and Optimization.
   - We care about finding the ”best” classifier, ”least” loss, ”maximizing” reward

2. Solving an Unconstrained Optimization Problem

   - We want to minimize $f(x)$
   - We start with $x_0$ (arbitrary choice), then for $t = 0, 1, 2, ..., T$ we update $x_{t+1} = x_t + d$, where $d$ is the direction.
   - $d = -\alpha f'(x)$, $\alpha$ = STEP SIZE, Gradient Descent converges to local minima
   - Convex function: Functions in which local minima $\equiv$ global minima
   - Taylor Series $f(x + \eta d) = f(x) + \eta d f'(x) + \frac{\eta^2 d^2}{2} f"(x) + ...$
   - For higher dimensions derivative becomes gradient
   - Newtons method update rule $x_{n+1} = x_n - \frac{f'(x_n)}{f"(x_n)}$, For higher dimension requires computing Hessian Matrix, If it is not invertible then this method cannot be applied. This method may not converge and would either enter infinite cycle or converge to saddle point instead of minima. It takes more time per iteration, is more computationally intensive and memory intensive.

# 9 Week 9

1. Constrained Optimization

   - Minimize $f(x)$ such that $g(x) \leq 0$
   - To check if any $x^*$ is a feasible solution we check $g(x^*) \leq 0$ and NO ”descent direction” should be a ”feasible direction”.
   - Descent direction: Any direction that reduces our functions value, d is a descent direction if $d^T \nabla f(x^*) < 0$.
   - Feasible direction: Any direction that takes to a point which satisfies all constraints, d is a feasible direction if $d^T \nabla g(x^*) < 0$.
   - Necessary condition for optimal solution: $\nabla f(x^*) = -\lambda \nabla g(x^*)$, $\lambda$ is positive
   - In equality case $\lambda$ can be negative or positive.

2. Convexity

   - A set $S \subseteq R^d$ is a convex set if $\forall x_1, x_2 \epsilon S$ then $\lambda x_1 + (1 - \lambda)x_2 \epsilon S$
   - Intersection of convex sets is also a convex set.
   - $z \epsilon R^d = \sum \lambda_i x_i$ is a convex combination of points in $S$ if $\lambda_i \geq 0$ and $\sum \lambda = 1$.
     The set of all such combinations is called Convex Hull(S)
   - Euclidean Balls in $R^d$: $\{x : ||x||_2 \leq \theta\}$ where $||x||_2 = \sqrt{\sum x_i^2}$

3. Convex functions

   - $f : r^d \rightarrow R$, $R^d$: any convex set, define epi(f) = $[xz]\epsilon R^{d+1}$ where $z \geq f(x)$.
     f is a convex function if epi(f) is a convex set.
   - f is a convex function iff $\forall x_1, x_2 \epsilon R^d$ and all $\lambda \epsilon [0, 1]$
     $f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)$
   - Assuming f is differentiable the f is convex iff
     $f(y) \geq f(x) + (y - x)^T \nabla f(x)$
   - If f is twice differentiable, $H \epsilon R^{dxd}$, $H_{ij} = \frac{\delta f}{\delta x_i \delta x_j}$
     f is convex iff H is positive semi definite matrix; eigenvalue(H)$\geq 0$
   - If f is a convex function, then all local minima of f are also global minima

# 10 Week 10

1. Properties of convex functions

   - If f and g are both convex then f + g is also convex
   - If f is convex and non decreasing and g is convex then f(g()) is also convex
   - If f is convex and g is linear then fog is convex
   - In general if f and g are convex then fog may not be convex.

2. Analytical Solution for Linear Regression: $w = (X^T X)^-1(X^T y)$

3. Constrained Optimization

   - minimize $f(x)$ such that $h(x) \leq 0$
   - Lagrangian function $L(x, \lambda) = f(x) + \lambda h(x)$, where $\lambda$ is a scalar.
     For $h(x) \leq 0$, the $\max_{\lambda \geq 0} L(x, \lambda) = f(x)$ with $\lambda = 0$
     For $h(x) > 0$, the $\max_{\lambda \geq 0} L(x, \lambda) = \infty$ with $\lambda = \infty$
   - $\min_x f(x) = \min_x \max_{\lambda \geq 0} L(x, \lambda)$, the DUAL would be $\max_{\lambda \geq 0} \min_x L(x, \lambda)$, where $\min_x L(x, \lambda)$ is an unconstrained problem.
   - $g(\lambda) = \min_x f(x) + \lambda h(x)$ is a convex

4. Relation between PRIMAL and DUAL

   - $g(\lambda^*) \leq f(x^*)$, value at DUAL optimum $\leq$ value at PRIMAL optimum (WEAK DUALITY)
   - if f and h are convex then STRONG DUALITY holds
   - KKT conditions for constrained optimization
     $\nabla f(x^*) + \lambda^* \nabla h(x^*) = 0$ Stationary condition
     $\lambda^* h(x^*) = 0$ Complimentary Slackness condition
     $h(x^*) \leq 0$ PRIMAL feasibility
     $\lambda^* \geq 0$ DUAL feasibility
     In general if $(x^*, \lambda^*)$ satisfies above conditions $\implies$ Local Optima

5. Support Vector Machine

   - min $\frac{1}{2}||w||^2$ such that $w^T x_i y_i \geq 1$
   - Data set: $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$