Rishabh Indoria

Machine Learning Foundations

August 5, 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 $f : R^d \to R^{d'}$ Decoder: $f : R^{d'} \to R^d$
   - 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
   - $P : R^d \to R_+$ 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
   - $f : R \to R$ is continuous if $\lim_{x \to x^*} f(x) = f(x^*)$
   - Differentiable if $\lim_{x \to x^*} \frac{f(x)-f(x^*)}{x-x^*} = f(x')$ 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 $\rightarrow$ Too small λ
  - Underfitting $\rightarrow$ Too large λ

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
   - $U^*U = I$
   - $\|Ux\| = \|x\|$
   - $U^{-1} = U^*$
   - $|\lambda| = 1$, where λ 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 $w_2 = X_2 - \frac{w_1 \cdot X_2}{\|w_1\|^2} w_1$

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 $D = \begin{bmatrix} \sigma_1 & 0 & 0 & ... & 0 \\ 0 & \sigma_2 & 0 & ... & 0 \\ 0 & 0 & 0 & ... & \sigma_r \end{bmatrix}$
   - σ_i are called singular values and $\sigma_i = \sqrt{\lambda_i}$
     where λ_i are eigenvalues of A^TA and $x_i are eigenvectors$
   - 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