SMAI-M19-01: Mathematical Foundations of ML - II

C. V. Jawahar

IIIT Hyderabad

August 1, 2019

Problem Space: Recap

- Example: Email classification
- Representation/Features
- Representation as vector in d dimension
- Lines, Planes and Hyper planes.
- Problem of Classification and Regression

Problem Space

- Problem of learning as finding a parameterized function.
- Role of w.
- Notion of "Training" and "Testing".
- *Typical Experimental Protocol - I*

Vectors

- Norm
- Products
- more

Matrices

- Dimensions, Addition, Multiplication
- Inverse and Trsnspose
- Special matrices
- Representation of a system of linear equations
- Determinant, Rank, Linearly independent rows.
- Linear Transformations
- Dimensionality Reduction

Matrices

- Introduction to Eigen Values and Eigen Vectors
- Read Chapters 1, 2 and 3. (4 is also familiar).

SMAI-M19-03: Mathematical Foundations of ML - III

C. V. Jawahar

IIIT Hyderabad

August 5, 2019

Problem Space-III

- Data: Training and Testing
- Goal: Learn a function $f(\mathrm{w}, \mathrm{x})$
- Can $y_i = f(\mathrm{w}, \mathrm{x}_i)$ for all i?
- Optimization problem, loss functions
- Clasification and Regression
- Comments on convex and non-convex optimization

Eigen Values and Eigen Vectors

- Recap: Types of matrices, Linear Transforms, notion of basis, Vector space etc.
- $Ax = \lambda x$; Numerical computation.
- Diagonal matrices, PD and PSD.
- Properties: Derminant, Trace etc. Recap.
- Eigen Decomposition

SVD

- SVD $A = UDV^T$
- Properties of these matrices
- Relationships with eigen values and vectors.
- Example of utility.

Low Rank in Data

- Matrices, Rank and Low-Rank Matrices
- Why Low-Rank matrices in real world data?
- Sub-spaces
- Data Matrix (noisy and noise-free)
- Low-rank Approximations

Dimensionality Reduction

- Curse of Dimensionality
- Linear Dimensionality Reduction
- Non-Linear Dimensionality Reduction
- Examples.
- Ref: Chapters 1-4 of the text book.

SMAI-M19-04: Mathematical Foundations of ML - IV

C. V. Jawahar

IIIT Hyderabad

August 8, 2019

Problem Space

- Problem of Learning Function
- Data that gets split into Train and Test.
- Desirability gets modelled as "Loss Function"
- Problem of Optimization.
- Optimization over what?
- Is LUT learning?
- The notion of "Overfitting"
- The notion of Generalization
- Occam's Razor

SVD

- $A = UDV^T$
- Properties and dimensionalities
- Relationship with eigen values and vectors
- Inverse of a matrix
- Rank-k approximation
- Data compression and Dimensionality Reduction
- Applications

Terms to Revise: Probability

- Random Variables
- Discrete and Continous Random Variables
- Expectation, Mean and Variance
- Probability Mass Function, Probability Density Function, Cumulative Distribution Function
- Uniform Density; Normal Density
- *Univariate and Multivariate Gaussian*
- *Bayes Theorem*
- *Read and Refresh: https://www.dropbox.com/s/2bmzs4dd2o82iha/randomvariables.pdf?dl=0*

Multivariate Normal Distribution

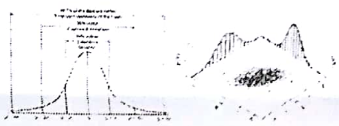
Univariate Normal Distribution $\mathcal{N}(\mu, \sigma)$

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Multivariate Normal Distribution $\mathcal{N}(\mu, \Sigma)$ with $x \in R^d$

$$p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{-\frac{1}{2}}} e^{-\frac{1}{2}((x-\mu)^T\Sigma^{-1}(x-\mu))}$$

Examples

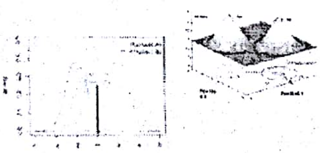


Bayes Theorem

$$\underbrace{p(x\,|\,y)}_{\text{posterior}} = \frac{\overbrace{p(y\,|\,x)}^{\text{likelihood}}\ \overbrace{p(x)}^{\text{prior}}}{\underbrace{p(y)}_{\text{evidence}}}$$

Maximum Liklihood and Maximum A Posteriori (MAP) classification

Optimal Bayesian Classifier



SMAI-M19-05: Mathematical Foundations of ML - V

C. V. Jawahar

IIIT Hyderabad

August 19, 2019

Review

- Problem of Model Selection
- Problem of Overfitting, and Regularization
- Bayes Theorem, Prior probability and posterior probability.
- Normal/Gaussian Distribution/Assumption

$$P(\omega_i/x) = \frac{p(x/\omega_i).P(\omega_i)}{p(x)}$$

$$P(\omega_i/x) = \frac{p(x/\omega_i).P(\omega_i)}{\Sigma_{j=1}^{c} p(x/\omega_j).P(\omega_j)}$$

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

Problem Space: Error Rate

- What is the error rate that we are interested in?
  1. Training
  2. True (or Test)
- What can we do with the error rate?
  - Performance analysis
  - Model Selection
- Can we estimate the true error?
- Notion of "validation" data.
- How good is this estimate?

Bayesian Optimal Classifier

- Discriminant Functionnd and Decision Boundary
- We decide a sample as in ω_1 if $P(\omega_1|x) > P(\omega_2|x)$

$$g(x) = P(\omega_1|x) - P(\omega_2|x) = \ln \frac{p(x|\omega_1)}{p((x|\omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

Special cases in Univariate

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} exp \left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right]$$

and multivariate

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp \left[-\frac{1}{2}[x-\mu]^T \Sigma^{-1} [x-\mu] \right]$$

Example

- $\Sigma_i = \sigma^2 I$
- $\Sigma_i = \Sigma$
- Σ_i

Parameter Estimation

- Samples and IID assumption. What is IID?
  - Independent:Each example is sampled independently from the others.
  - Identically Distributed: All examples are sampled from the same distribution
- Learning parameters of the distributions from samples.
- Bayesian Estimation
  - Assumes parameters are random variables with some known prior distribution
  - Observing examples turns prior distribution over parameters into aposterior distribution
- MAP and ML Estimations

$$\theta_i^* = \arg\max_{\theta_i} p(\theta_i|\mathcal{D}_i, y_i) = \arg\max_{\theta_i} p(\mathcal{D}_i, y_i|\theta_i)p(\theta_i)$$

$$\theta_i^* = \arg\max_{\theta_i} p(\mathcal{D}_i, y_i|\theta_i)$$

- Example: Gaussian (next lecture)

SMAI-M19-06: Mathematical Foundations of ML - VI

C. V. Jawahar

IIIT Hyderabad

August 22, 2019

Where are we?

- Three different views of Classification:
  1. Generative/Probabilistic
  2. Discriminative/Decision boundary
  3. Distance based/Nearest neighbour
- Bayes Theorem and Bayesian Optimal Classification
- $\Sigma_i = \sigma^2 I$, $\Sigma_i = \Sigma$, Σ_i
- Mahalanobis Distance

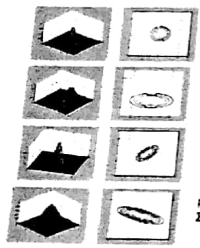
Multivariate Gaussians

Univariate

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

and multivariate

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp\left[-\frac{1}{2}[x-\mu]^T\Sigma^{-1}[x-\mu]\right]$$



$\mu = [0, 0]$ $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

$\mu = [0, -2]$ $\Sigma = \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}$

$\mu = [0, 0]$ $\Sigma = \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}$

$\mu = [0, 0]$ $\Sigma = \begin{bmatrix} 5 & -1.5 \\ -1.5 & 1 \end{bmatrix}$

Parameter Estimation

Rohit has captained India in 10 matches and 8 times india won. What is the probability that India will win if Rohit is made a captain in the next match?
Ans: $\frac{8}{10} = 0.8$
If number of wins follows a binomial distribution with parameter θ,

$$P(k \ wins|n \ matches) =_n C_k \cdot \theta^k \cdot (1-\theta)^{n-k}$$

Let us compute the probability with $\theta = 0.1, 0.2 etc.$

0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9

When will this be maximum?
8 wins in 10 matches is highly likely when $\theta = 0.8$.

$$P(k \ wins|n \ matches) =_n C_k \cdot \theta^k \cdot (1-\theta)^{n-k}$$

$$\frac{dp(\mathcal{D}|\theta)}{d\theta} = nC_k\left[k\theta^{k-1}(1-\theta)^{n-k} - (n-k)\theta^k(1-\theta)^{n-k-1}\right]$$

$$= nC_k\left[\theta^{k-1}(1-\theta)^{n-k-1}\right](k(1-\theta)-(n-k)\theta)) = 0$$

$$\theta = \frac{k}{n}$$

Use of Prior probability

Maximum Log Likelihood

$$\theta^* = \arg\max_\theta p(\mathcal{D}|\theta) = \arg\max_{\theta_n} \Pi_{j=1} p(x_j|\theta)$$

$$\theta^* = \arg\max_\theta \ln p(\mathcal{D}|\theta) = \arg\max_{\theta_n} \sum_j lnp(x_j|\theta)$$

$$\nabla_{\theta_n} \sum_{j=1} (x_j|\theta) = 0$$

$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

$$p(x) = \frac{1}{(2\pi)^{d/2}|\sigma|^{1/2}} exp\left[-\frac{1}{2}[x-\mu]^T\Sigma^{-1}[x-\mu]\right]$$

Parameter Estimation

- Samples and IID assumption. What is IID?
  - **Independent:**Each example is sampled independently from the others.
  - **Identically Distributed:** All examples are sampled from the same distribution
- Learning parameters of the distributions from samples.
- Bayesian Estimation
  - Assumes parameters are random variables with some known prior distribution
  - Observing examples turns prior distribution over parameters into aposterior distribution
- MAP and ML Estimations

$$\theta_i^* = \arg\max_{\theta_i} p(\theta_i|\mathcal{D}_i, y_i) = \arg\max_{\theta_i} p(\mathcal{D}_i, y_i|\theta_i)p(\theta_i)$$

$$\theta_i^* = \arg\max_{\theta_i} p(\mathcal{D}_i, y_i|\theta_i)$$

- **Example:** Gaussian

SMAI-M19-07: Linear Models: Regression

C. V. Jawahar

IIIT Hyderabad

August 26, 2019

Review

- Problem Setting and Formulation.
- Basic maths and view points:
  - Linear Algebra, Geometric view
  - Probability, Bayesian View
  - Distance based and NN methods
- Classification, Regression and Structured Prediction
- Linear and Non-Linear Models
- Convex and Non-Convex Optimization

Problem of Linear Regression/MSE

Model: $y_i = \mathbf{w}^T\mathbf{x} + \epsilon_i$

$$\min_{\mathbf{w}} \sum_{i=1}^{N} \epsilon_i^2 = \min_{\mathbf{w}} \sum_{i=1}^{N} (y_i - \mathbf{w}^T\mathbf{x}_i)^2$$

$$\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} - \begin{bmatrix} \mathbf{x_1}^T \\ \mathbf{x_2}^T \\ \vdots \\ \mathbf{x_N}^T \end{bmatrix} \mathbf{w} \tag{1}$$

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Regression and MSE: Probabilistic View

$$y_i = \mathbf{w}^T\mathbf{x} + \epsilon_i$$

$$\epsilon = \mathcal{N}(0, \sigma^2)$$

$$p(y_i|\mathbf{x}_i, \mathbf{w}) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(y_i - \mathbf{w}^T\mathbf{x})^2}{2\sigma^2}\right)$$

$$L(\mathbf{w}) = L(\mathbf{w}, \mathbf{X}, \mathbf{Y}) = \prod_{i=1}^{N} p(y_i|\mathbf{x}_i, \mathbf{w})$$

$$l(\mathbf{w}) = \log L(\mathbf{w}) = \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(y_i - \mathbf{w}^T\mathbf{x})^2}{2\sigma^2}\right)$$

$$= K\frac{1}{2}\sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x})^2$$

$$\Rightarrow \min_{\mathbf{w}} \sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x}_i)^2$$

Regularization

- Bias and Variance View
- Regularization
- Ridge Regression
- Lasso

Challenges in Big Data Settings

- When N is large
  1. Iterative solution scheme → "Gradient Discent"
- When N is very large?
  1. Stochastic Versions
  2. Work on subsets
- When all the data can not be stored/available
  1. Online variants