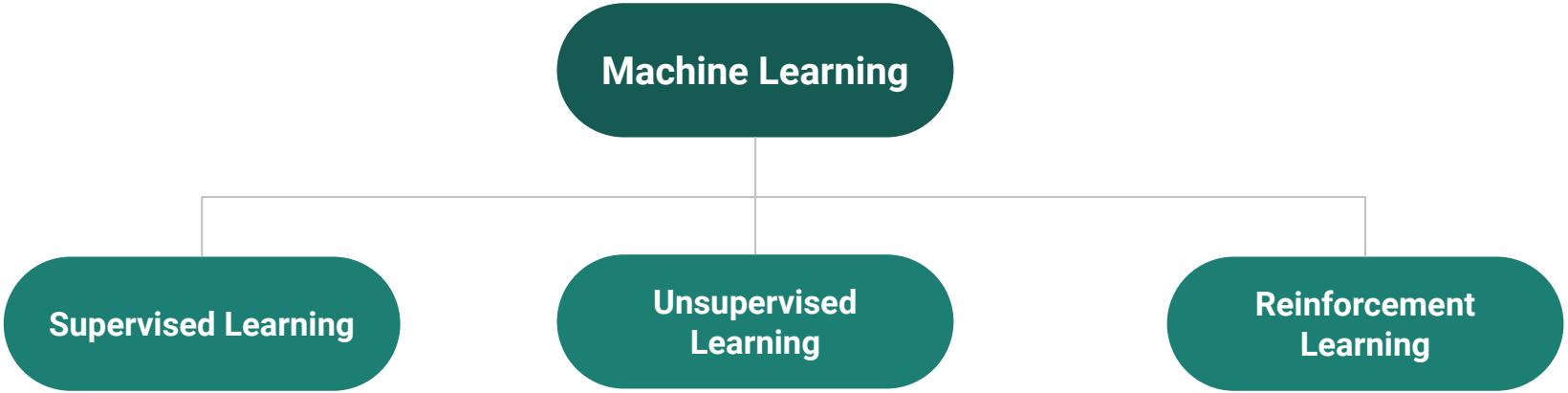
CSE 575: Statistical Machine Learning (Spring 2021)
Instructor: Nupur Thakur

Midterm Review

Table of contents

Types of Machine Learning



Bias $= E[\hat{\theta}] - \theta = 0$ $\quad E[\hat{\theta}] = \theta$ — estimated parameter is unbiased

Maximum Likelihood Estimation

$E[\hat{\mu}] = \mu \Rightarrow$ Unbiased $\quad E[x] = \int x\, p(x)\, dx$

$\neq \mu \Rightarrow$ biased.

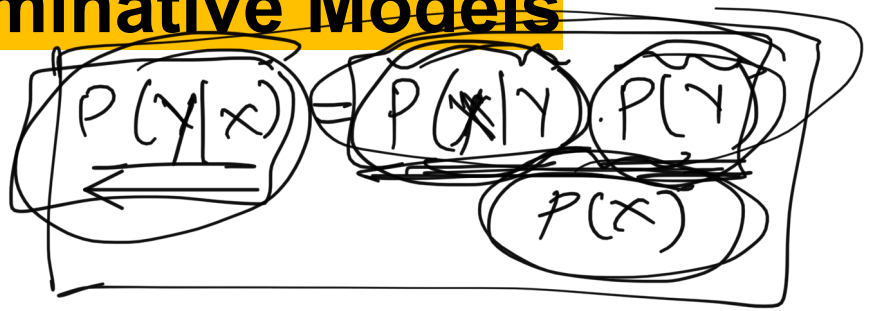
- Given some training data and assuming a parametric model $p(x|\theta)$; what specific θ will fit/explain the data best?
- To consider all the samples denoted by $D=\{x_1, x_2, \ldots\ldots, x_n\}$, assume that all the samples are i.i.d - independent and identically distributed.
- So, data likelihood represented by L(θ) is -

$$L(\boldsymbol{\theta}) = P(D|\boldsymbol{\theta}) = \prod_i P(x_i|\theta)$$

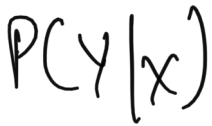
$$\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}}\, p(\mathbf{x}|\boldsymbol{\theta})$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Generative vs Discriminative Models



- Generative models -
  - Learn P(y) and P(x|y).
  - Ex: Bayesian classifier, Naive Bayes.
- Discriminative models -
  - Directly learn P(y|x)
  - Ex: Logistic Regression



Naive Bayes

- The "naive" conditional independence assumption: each feature is (conditionally) independent of every other feature, given the label, i.e., $p(x_i \mid \{x_j \text{ for any } j \neq i\}, y) = p(x_i \mid y)$
- The predicted label is given by -

$$\hat{y} = \underset{y}{\operatorname{argmax}} P(y) \prod_{i=1}^{d} p(x_i \mid y)$$

Linear Regression

- Regression - A training set of n samples $<x^{(i)}, y^{(i)}>$ where $y^{(i)}$ is a continuous “label” (or target value) for $x^{(i)}$
- Linear regression - modeling the relation between y and x via a linear function $y \approx w_0 + w_1 x_1 + \ldots + w_d x_d = \mathbf{w}^t \mathbf{x}$
- The error is given as - $||e||^2 = ||y - Xw||^2$

Logistic Regression

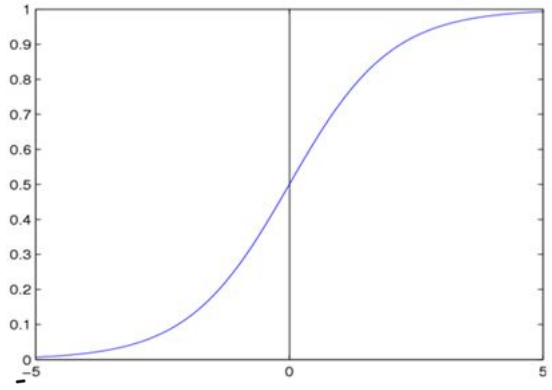
$P(Y|X) = \text{logistic function}_{w^T x}$

$w^T x = 0$

- Training set: n labelled samples <x(i), y(i)>
- Use the logistic function for modeling P(y|x), considering only the case of $y \in \{0,1\}$

$$P(y=0|\mathbf{x}) = \frac{1}{1+\exp\left(w_0 + \sum_{i=1}^{d} w_i x_i\right)}$$

$$P(y=1|\mathbf{x}) = \frac{\exp\left(w_0 + \sum_{i=1}^{d} w_i x_i\right)}{1+\exp\left(w_0 + \sum_{i=1}^{d} w_i x_i\right)}$$



$$\sigma(t) = \frac{1}{1+e^{-t}} = \frac{e^t}{1+e^t}$$

Gradient ascent

$$w_{t+1} = w_t + \alpha \frac{\partial \bar{L}(w)}{\partial w}$$

(updated, previous, loss)

Support Vector Machines (SVM)

- Key idea - To find the decision boundary such that the margin is maximized.
- Data - $\langle x^{(i)}, y^{(i)} \rangle$, $y^{(i)} \in \{-1,1\}$, $x^{(i)} \in \boldsymbol{R}^d$, for all i=1,…,n
- Plane equations-

$$\left.\begin{aligned} w^T x + b &= 1 \\ w^T x + b &= -1 \end{aligned}\right\} \text{ – Margin hyperplanes.}$$

$$w^T x + b = 0 \quad \text{— Decision boundary.}$$

- Margin -

$$d = \frac{2}{\|w\|}$$

SVM - Problem Formulation

$$\{\mathbf{w}^*, b^*\} = \underset{\mathbf{w},b}{\operatorname{argmin}} \|\mathbf{w}\| \quad or \quad \{\mathbf{w}^*, b^*\} = \underset{\mathbf{w},b}{\operatorname{argmin}} \tfrac{1}{2}\|\mathbf{w}\|^2$$

Subject to

$$\begin{cases} \mathbf{w}^t\mathbf{x}^{(i)} + b \geq 1 & \text{for } y^{(i)} = +1 \\ \mathbf{w}^t\mathbf{x}^{(i)} + b \leq -1 & \text{for } y^{(i)} = -1 \end{cases}$$

The constraints can be combined into:

$$y^{(i)}(\mathbf{w}^t\mathbf{x}^{(i)} + b) - 1 \geq 0 \quad \forall i$$

Soft-Margin SVM Formulation

$$\{\mathbf{w}^*, b^*\} = \underset{\mathbf{w},b}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{w}\|^2 + C\left(\sum_i \xi_i\right)$$

Parameter controls the penalty.

Slack variable

Very high values → Hard margin SVM

Very low values → Larger misclassifications

subject to

$$w^t x^{(i)} + b \geq 1 - \xi_i \quad \text{for } y^{(i)} = +1$$

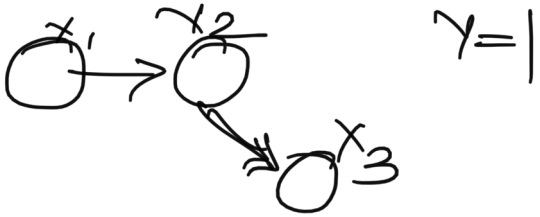
$$w^t x^{(i)} + b \leq -1 + \xi_i \quad \text{for } y^{(i)} = -1$$

$$\xi_i \geq 0, \forall i$$

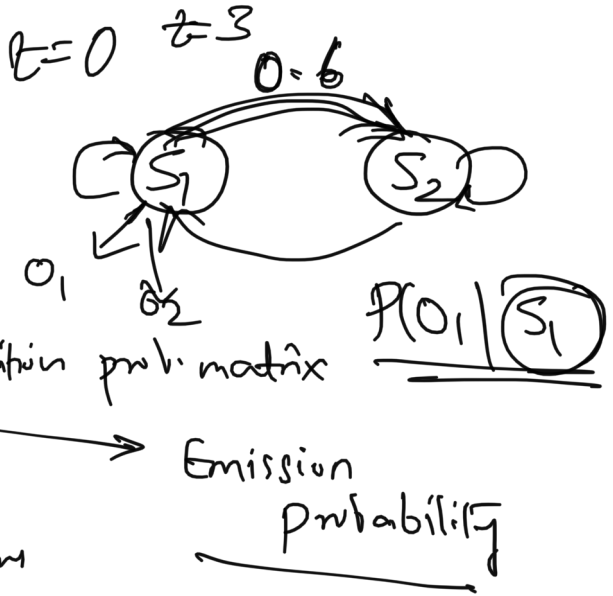
ϕ

$\langle x_i, x_j \rangle$

Hidden Markov Model (HMM)



- Dynamic Bayesian network - modeling the process indexed by time.
- Two assumptions -
  - First-order Markov chain - $P(s^t=S_j \mid s^{t-1}=S_i)$
  - $a_{ij} = P(s^t=S_j \mid s^{t-1}=S_i)$, $1 \le i,j \le N$, for any t
- Specifying HMM - Θ, **Ω**, π, A, B



Problems in HMM

- For a given HMM $\Lambda=\{\Theta,\Omega,A,B,\pi\}$
  - Estimation of model parameters
  - Given an observation sequence **O**= $\{o^1,o^2, \dots ,o^k\}$, what is the most likely state sequence S = $\{s^1, s^2, \dots ,s^k\}$ that has produced O?
  - How likely is an observation O?

Trained HMM

Decoding

P(O)

HMM Parameter Estimation

- Given labeled data - state and observation

A — $$t(S_i|S_j) = \frac{\text{number of } (s^t = S_i, s^{t-1} = S_j)}{\text{number of } S_j}$$

B — $$e(o_r|S_j) = \frac{\text{number of } (o^t = o_r, s^t = S_j)}{\text{number of } S_j}$$

- Observation sequence — Forward-Backward Algo.

HMM State Estimation

- Given an observation (sequence) O= {o_1, o_2, … ,o_k}, what is the most likely state sequence S={s_1, s_2, … ,s_k} that has produced O?

Vitubi

Initialization $\delta_{S_i}(1) = t(S_i|s^*)e(o^1|S_i), \quad \forall S_i \in \Theta$

Induction:
for 2≤*t*≤*k*, do

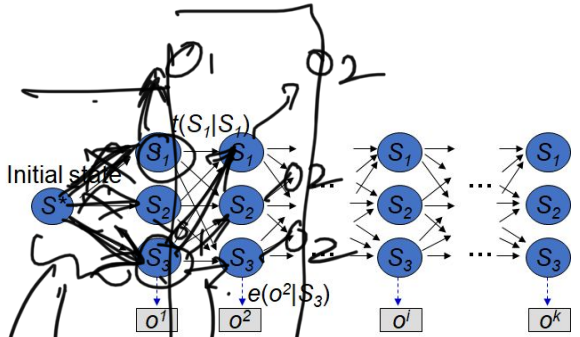
$$\delta_{S_i}(t) = \max_{S_j} t(S_i|S_j)e(o^t|S_i)\delta_{S_j}(t-1)$$

$$\psi_{S_i}(t) = \operatorname{argmax}_{S_j} t(S_i|S_j)e(o^t|S_i)\delta_{S_j}(t-1)$$

Termination:

- The probability of the best state sequence: $\max_{S_j} \delta_{S_j}(k)$
- The best last state: $\hat{s}^k = \operatorname{argmax}_{S_j} \delta_{S_j}(k)$
- Back trace to get other states:

$$\hat{s}^t = \psi_{\hat{s}^{t+1}}(t), \text{ for } t = k-1, \ldots, 1.$$



S^4, S_3, S_1

HMM - Evaluate P(O)

- How likely is an observation O?

$$P(O) = \sum_{S} P(S, O)$$

- Forward algorithm-

Initialization: $\alpha_{S_i}(1) = t(S_i|s^*)e(o^1|S_i), \quad \forall S_i \in \Theta$

Induction:
for $2 \leq t \leq k$, do $\alpha_{S_i}(t) = \sum_{S_j} t(S_i|S_j)e(o^t|S_i)\alpha_{S_j}(t-1)$

Termination: $P(O) = \sum_{S_j} \alpha_{S_j}(k)$

Sample Questions

Problem 1:

The following data is used for training Naive Bayes binary classifier. The last column is the binary class label; Each of the first 4 columns is a binary feature, and each row is a training example.

X1	X2	X3	X4	Y
1	0	0	0	1
0	1	1	0	1
1	0	0	1	0
0	1	1	1	0
1	1	1	0	1

Problem 2:

State true/false along with justification:

1. Logistic Regression may give us a non-linear classifier, depending on how the training examples are distributed in the feature space.

Ans – False because logistic regression is a linear classifier, producing a linear decision boundary.

Problem 3:

Multiple Choice Questions:

Which of the following statements is true about HMM?

a. Given some initial state and transition matrix, it is possible that there exists a state that you can never achieve.
b. The elements of a column of the state transition matrix always sum to 1.
c. The sum of a row of the transition matrix may not be 1.

Ans - (a)