Midterm practice

Summaer Quarter

Duration: 1 hours 45 minutes

Name: DU ID:

1. **This is closed book/notes exams**
2. **Please write your name and DU ID before starting the exam.**
3. **Show all the step of your answer and justify you answer/steps**

Problem 1.(.5 points each.)

1a. What is the difference in supervised and unsupervised machine learning.

supervised machine learning both feature x_i and label y_i are known for the training. In unsupervised setting y_i are not known.

1b. Why are generative model called generative and discriminative model discriminative?

Generative method models $P(X|y=c)$, class conditional densities. In fact one can generate new data

In discriminative method directly model $P(y=c|x)$. No generative capacity.

1c. Given some observation D write the M.L.E formualtion of estimation of parameters θ and MAP estimation of parameters θ.

$\theta^{MLE} = \arg\max_{\theta} P(D|\theta)$, $\theta^{MAP} = \arg\max_{\theta} P(\theta|D)$

1d. What is the set of values poisson random variable takes(called support).

set of non negative integers, $\{0, 1, 2, \cdots\}$

1e. Does strictly convex function has unique global minumum.(yes/no).

1f. Conditional independence means

$$P(X, Y|Z) = P(X|Z)\, P(Y|Z)$$

1g. In linear regression, which norm does feature selection(ℓ_1 or ℓ_2)

ℓ_1

Problem 2.(2+2+1+.5 points.)

2a. Let $\boldsymbol{x} \in \{1, \cdots, K\}^D$, i.e $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$ and $x_i \in \{1, \cdots, K\}$. In generative model we need to specify class conditional distribution $P(\boldsymbol{x}|y = c)$. If we don't assume conditional independence on features, given class label how many parameters we need to estimate.

$C(K^D - 1)$, where C is total number of classes

2b. If we assume conditional independence on features given class label, how may parameters we need to estimate.

$C(K-1)D$

2c. Assuming conditional independence on feature given class label leads to Naive Bayes classifier. Write right hand side of following equation for naive bayes classifier.

$p(\boldsymbol{x}|y = c, \boldsymbol{\theta}) = \prod_{i=1} P(x_i|y=c, \theta_{ci})$

2d. If you have less data, which model(model in 2a or 2b(naive Bayes)) is likely to give you less test set error. Explain in no more than 1(preferred) or 2 line.

2a. with less data Naive Bayes (less parameters) will not overfit. Parameters will be more stable reliable

Problem 3.(4= (2+2) points.) Let scalar x be drawn from $\mathcal{N}(\mu_i, \sigma^2) = \frac{1}{\sqrt{(2\pi)}\sigma} \exp(\frac{(x-\mu)^2}{-2\sigma^2})$ (1-d Gaussian distribution). If we have N, I.I.D samples $\mathcal{D} = \{(x_i)\}_{i=1}^{i=N}$, then compute the MLE estimate of μ and σ. look into book for Gaussian MLE estimation.

Problem 4(2 points) In linear regression $y = \boldsymbol{w}^T\boldsymbol{x} + \epsilon$ estimate of $\boldsymbol{w}$ is $\hat{\boldsymbol{w}} = (X^TX)^{-1}X^Ty)$. Hence residual vector e against fitted line is $\boldsymbol{e} = y - X\hat{\boldsymbol{w}}$. Show that residual vector is orthogonal to columns of X. Note X contains observation $\boldsymbol{x}_i$ along rows.

If e is orthogonal to the columns of X then inner product (dot product) of e and columns of X will be zero.

Hence

$e^T X$ (Note that this expression will take the dot product of e and each column of X. Hence result will be a zero row vector)

$= (y - X\hat{w})^T X$

$= (y^T - \hat{w}^T X^T) X$

$= y^T X - \left[(X^TX)^{-1} X^T y\right]^T X^T X$ (substituting value of $\hat{w}$)

$= y^T X - (X^T y)^T \left[(X^T X)^{-1}\right]^T X^T X$

$= y^T X - y^T X \underbrace{\left((X^T X)(X^T X)^{-1}\right)^T}_{= I}$ (Note $(X^TX)^T = X^TX$, I = identity matrix)

$= y^T X - y^T X = \mathbf{0}^T$

$$\mathbf{0} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ \vdots \end{bmatrix}_{d \times 1}$$ if feature are d dimensitional

note that if $y_{N \times 1}$, $X_{N \times d}$ then $\underset{1 \times N}{y^T} \underset{N \times d}{X} \rightarrow 1 \times d$

Hence subtracting same $1 \times d$ vector will give $1 \times d$ zero vector