Solution Key

Midterm

Machine learning Summer 2018, Practice test

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**
4. **Please write clearly and upto the point.**

Problem 1.(12 =2+2+2+6 points.)

1a. What is the difference in supervised and unsupervised machine learning. In supervised setting feature x_i and label y_i is known. In unsupervised only feature x_i of data is known to find interesting Patterns

1b. Why are generative model called generative and discriminative model discriminative? - generative method models $P(X|y=c)$, class conditional density. Infact one can generate data too.
- In discriminative method directly model $P(y=c|X)$. No generative capacity

1c. Given some observation $\mathcal{D}$ write the M.L.E formualtion of estimation of paramters θ and MAP estimation of parameters θ.

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

1d. For a probalility mass function p Entropy(measure of uncertainity) is given by $\mathcal{H}(X) = -\sum_{k=1}^{K} p_k \log p_k$. One way to measure the dissimilarity of two probability distributions, p and q, is known as the Kullback-Leibler divergence(KL) or relative entropy.(Note that this is not a distance or metric as it is not symmetric). It is defined as $\mathcal{KL}(p|q) \sum_{k=1}^{k=K} p_k \log \frac{p_k}{q_k}$. Show that

$$\mathcal{KL}(p|q) = -\mathcal{H}(p) + \mathcal{H}(p,q) \text{ where } \mathcal{H}(p,q) = \sum_{k=1}^{k=K} p_k \log q_k$$

is called cross entropy. Follow the definitions

Problem 2.(14=2+4+4+4 points.) Write right hand side of following.

2a. Conditional independent means

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

2b. Let $\boldsymbol{x} \in \{1, \cdots, K\}^D$ where K is the number of values for each feature. In generative model we need to specify class condition 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. $C(2^D-1)$ where C is total no of classes.

2c. If we assume conditional independence on features given class label, how may parameters we need to estimate. $C(D-1)$

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

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

Problem 3.(10= (5+5) points.) Let scalar $x \sim \mathcal{N}(\mu_i, \sigma^2) = \frac{1}{\sqrt{(2\pi)\sigma}} \exp(\frac{(x-\mu)^2}{-2\sigma^2})$ (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 estimate.

Problem 4.(5 = 2+3 points.) In the Bayesian approach to decision theory, the optimal action, having observed x, is defined as the action $\hat{y}$ that minimizes the posterior expected loss $\sum_y L(y, \hat{y}) p(y|x)$. $\mathbf{0-1}$ loss is defined as $L(y, \hat{y}) = 0$ if $y = \hat{y}$ and $L(y, \hat{y}) = 1$ if $y \neq \hat{y}$. Calculate the posterior expected loss and prove that the action that minimizes the expected loss is the posterior mode or MAP estimate argmax $p(y|x)$.

look into book

Problem 5.(10 = 4+6 points.) Write the model specification for logistic regression and also compute negative log likelihood (NLL) given data $\mathcal{D}\{(x_i, y_i)\}_{i=1}^{i=N}$ where $x_i \in R^d$.

look into book, logistic regression