MIDTERM PRACTICE PROBLEMS

Friday, 4th May 2018

1 Decision Tree

The training set is given below. W, X, Y are the attributes $\in \{0, 1\}$ and Z $\in \{-1, +1\}$ is the class variable.

#	W	X	Y	Z
1	0	0	0	-1
2	0	0	1	-1
3	1	0	0	+1
4	0	1	0	-1
5	0	1	1	+1

(a) Which attribute has the highest information gain? Justify your answer with calculations.

(b) Draw the complete decision tree for this dataset using ID3 with information gain criterion.

(c) Given the validation set

#	W	X	Y	Z
6	1	1	0	+1
7	0	1	1	-1
8	1	1	1	-1

Can you construct a decision tree with 100% accuracy on validation set as well as training set. If yes then draw such a decision tree, and if no then explain why it is not possible.

2 Perceptron

Design (specify θ for) a two-dimensional input perceptron (with an additional bias or offset term) that computes the following boolean functions. Assume $T = 1$ and $F = -1$. If no perceptron exists, state why.

[AND]

x_1	x_2	y
-1	-1	-1
-1	$+1$	-1
$+1$	-1	-1
$+1$	$+1$	$+1$

[XOR]

x_1	x_2	y
-1	-1	-1
-1	$+1$	$+1$
$+1$	-1	$+1$
$+1$	$+1$	-1

3 Logistic Regression

Given data $\mathcal{D} : \{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ldots, (\boldsymbol{x_n}, y_n)\}$ and the query point $\boldsymbol{x}$ we would like to choose $\boldsymbol{w}$ that minimizes the loss which is the negative log likelihood weighted appropriately.

$$J(\boldsymbol{w}, \boldsymbol{x}) = -\sum_{i=1}^{n} \alpha_i(\boldsymbol{x})[y_i \log h_{\boldsymbol{w}}(\boldsymbol{x}_i) + (1 - y_i)\log(1 - h_{\boldsymbol{w}}(\boldsymbol{x}_i))]$$

where

$$h_{\boldsymbol{w}}(\boldsymbol{x}_i) = \frac{1}{1 + \exp(-\mathbf{w}^T \boldsymbol{x}_i)} \qquad \alpha_i(\boldsymbol{x}) = \exp\left(-\frac{||\boldsymbol{x} - \boldsymbol{x_i}||^2}{2\sigma}\right)$$

$\sigma > 0$ is a hyperparameter. Whenever we receive a new query point we must train the model with the new weights $\alpha_i(\boldsymbol{x})$.
From class you know that $\nabla_w J(\boldsymbol{w}) = \sum_{i=1}^{n}(h_{\boldsymbol{w}}(\mathbf{x}_i) - y_i)\boldsymbol{x}_i$ where $J(\boldsymbol{w})$ is the negative log likelihood for logistic regression.

(a) Given a test point $\boldsymbol{x}$, find the gradient of $J(\boldsymbol{w}, \boldsymbol{x})$ with respect to $\boldsymbol{w}$.

(b) Given a test point $\boldsymbol{x}$, find the Hessian of $J(\boldsymbol{w}, \boldsymbol{x})$ with respect to $\boldsymbol{w}$ and show that it is positive semi-definite.

(c) Given a test point $\boldsymbol{x}$, write the gradient descent update rule.

(d) Is this locally weighted regression a parametric or non parametric method?

4 Linear Regression

(a) Describe one application of linear regression. Please define clearly what are your input, output, and features.

(b) Given a dataset $\{(x^{(i)}, y^{(i)})\}_{i=1}^{M}$ in a two dimensional space. The objective function of linear regression with square loss is

$$J(w_1, w_2) = \frac{1}{2}\sum_{i=1}^{M}(y_i - (w_1 x_1^{(i)} + w_2 x_2^{(i)}))^2, \tag{1}$$

where w_1 and w_2 are feature weight to be learned. Write down one optimization procedure that can learn w_1 and w_2 from data. Please be as explicit as possible.

(c) Prove that Eq. (1) has a global optimal solution.

5 Maximum Likelihood Estimation

Let $\mathcal{D} : \{x_1, x_2, \ldots, x_n\}$ i.i.d and be drawn from the distribution

$$f(x|\lambda) = \lambda x^{-\lambda-1} \quad \text{where} \quad \lambda > 1\ , x \geq 1$$

(a) Write the likelihood function of λ.

(b) Find the maximum likelihood estimator of λ.

(c) Given that $\mathcal{D} : \{1, e, e^2, e^3\}$, compute the maximum likelihood estimate of λ using the previously found estimator. [Hint: Are you looking closely?]

6 Maximum Likelihood Estimation 2

We observe the following data consisting of four independent random variables $X_n, n \in \{1, \ldots, 4\}$ drawn from the same Bernoulli distribution with parameter θ (*i.e.*, $P(X_n = 1) = \theta$): $(X_1, X_2, X_3, X_4) = (1, 1, 0, 1)$.

(a) Give an expression for the log likelihood $l(\theta)$ as a function of θ given this specific dataset.

(b) Give an expression for the derivative of the log likelihood for this specific dataset.

(c) What is the maximum likelihood estimate $\hat{\theta}$ of θ?

7 Kernel

Given vectors $\mathbf{x}$ and $\mathbf{z}$ in $\mathbb{R}^2$, define the kernel $K_\beta(\mathbf{x}, \mathbf{z}) = (1 + \beta \mathbf{x} \cdot \mathbf{z})^3$ for any value $\beta > 0$. Find the corresponding feature map $\phi_\beta(\cdot)$. What are the similarities/differences from the kernel $K(\mathbf{x}, \mathbf{z}) = (1 + \mathbf{x} \cdot \mathbf{z})^3$, and what role does the parameter β play?