School of Computing and Information Systems
The University of Melbourne
COMP30027 MACHINE LEARNING (Semester 1, 2019)

Tutorial exercises: Week 7

1. What is the **Gradient Descent** method, and why is it important?

2. What is **Regression**? How is it similar to **Classification**, and how is it different?

   (a) What is **Linear Regression**? In what circumstances is it desirable, and it what circumstances is it undesirable?

   (b) How do we build a (linear) regression model? What is **RSS** and what advantages does it have over (some) alternatives?

3. Recall that the update rule for Gradient Descent with respect to RSS is as follows:

$$\beta_k^{i+1} \quad := \quad \beta_k^i + 2\alpha \sum_{j=1}^{N} x_{jk}(y_j - \hat{y_j^i})$$

   Build a Linear Regression model, using the following instances:

x	y
1	1
2	2
2	3

4. What is **Logistic Regression**?

   (a) How is Logistic Regression similar to **Naive Bayes** and how is it different? In what circumstances would the former be preferable, and in what circumstances would the latter?

   (b) What is "logistic"? What are we "regressing"?

   (c) How do we train a Logistic Regression model? In particular, what is the significance of the following:

$$\text{argmax}_\beta \sum_{i=1}^{n} y_i \log h_\beta(x_i) + (1 - y_i)\log(1 - h_\beta(x_i))$$

1. What is the **Gradient Descent** method, and why is it important?

Gradient Descent : is mechanism for finding MIN of multivariate function, where we can find its partial derivative.

use?

determine the regression weights which minimise an error function over some training data.

2. What is **Regression**? How is it similar to **Classification**, and how is it different?

   (a) What is **Linear Regression**? In what circumstances is it desirable, and it what circumstances is it undesirable?

   (b) How do we build a (linear) regression model? What is **RSS** and what advantages does it have over (some) alternatives?

(a) Linear Regression : build a linear model to predict target value. by finding a weight for each attribute. $\sum_i w_i a_i$

(b) By learning the weights using Gradient Descent, with respect to an $\underbrace{\text{error function}}_{\text{is RSS}}$. $\Rightarrow$ min

3. Recall that the update rule for Gradient Descent with respect to RSS is as follows:

$$\beta_k^{i+1} := \beta_k^i + 2\alpha \sum_{j=1}^{N} x_{jk}(y_j - \hat{y}_j^i)$$

Build a Linear Regression model, using the following instances:

x	y
1	1
2	2
2	3

Gradient Descent:

- iterately improves estimate of the prediction line (β)
- is based on partial deriates of the error function (RSS)
- tries to find a weight (β_k) for each attribute ($k \in [1 \dots p]$).

Step 1: $\hat{y} = \beta x = 0 + 0x$

init →

因为只有一个 feature x

Step 2: choose learning rate α

$\alpha = 0.05$.

Step 3: $\beta_k^1 = \beta_k + 2\alpha \sum_i x_{ik}(y_i - \hat{y}_i)$

we have 3 instances. begin with $\hat{y} = 0 + 0x$, so $\beta = \langle 0, 0 \rangle$

- for first point $(1,1)$ ⇒ $\hat{y}_1 = 0 + 0 \cdot 1 = 0$ But $y_1 = 1$ ⇒ error $= +1$
- for second point $(2,2)$ ⇒ $\hat{y}_2 = 0 + 0 \cdot 2 = 0$ But $y_2 = 2$ ⇒ error $= +2$
- for third point $(2,3)$ ⇒ $\hat{y}_3 = 0 + 0 \cdot 2 = 0$ But $y_3 = 3$ ⇒ error $= +3$

first update:

$$\beta_0' = \beta_0 + 2\alpha \sum_i x_{i0} (y_i - \hat{y}_i)$$

$$\beta_0' = 0 + 2 \times 0.05 \times [(1)(1-0) + (1)(2-0) + (1)(3-0)]$$

$$= 0.6$$

$$\beta_1' = \beta_1 + 2\alpha \sum_i x_{i1} (y_i - \hat{y}_i)$$

$$= 0 + 2 \times 0.05 [(1)(1-0) + (2)(2-0) + (2)(3-0)]$$

$$= 1.1$$

Now $\hat{y} = 1.1x + 0.6$

Repeat until $\beta_0^{(k)} = -0.5 \quad \beta_1^{(k)} = 1.5$

$$\hat{y} = 1.5x - 0.5$$

4. What is **Logistic Regression**?

(a) How is Logistic Regression similar to **Naive Bayes** and how is it different? In what circumstances would the former be preferable, and in what circumstances would the latter?

(b) What is "logistic"? What are we "regressing"?

(c) How do we train a Logistic Regression model? In particular, what is the significance of the following:

$$\text{argmax}_\beta \sum_{i=1}^{n} y_i \log h_\beta(x_i) + (1 - y_i) \log(1 - h_\beta(x_i))$$

4. Logistic Regression: We build a (linearly) regression model where the target is (close to) 1 for instance of the positive class. and (close to) 0 for instance of the negative class.

(a) Similarity:

- Both methods are attemping to find the class c for a test instance T by maximising $P(c|T)$

difference:

- in NB: we assume attributes are independence
- in Logistic Regression: without assumption, model directly

(b) logistic function $\frac{1}{1+e^{-m}}$ $m = (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots)$

has a range $[0, 1]$

(c)

β maximize objective function