Unsupervised Learning + Revision

COMP9417, 23T1

1. Unsupervised Learning
2. Revision
3. Gradient Descent Question

Section 1

Unsupervised Learning

Unsupervised Learning

Learning without any labels.

For example, – Cluster analysis (i.e grouping users of a social media, classifying similar events/data without knowing any other information) – Signal separation (i.e PCA, SVD)

Unsupervised Learning

Learning without any labels.

For example, – Cluster analysis (i.e grouping users of a social media, classifying similar events/data without knowing any other information) – Signal separation (i.e PCA, SVD)

The content this week is light, so I'll go straight to the lab to explain it.

Section 2

Revision

Identities

Of course you need to remember anything from first year/high school mathematics (i.e basis calculus, log laws, basic vector/matrix identities).

Identities

Of course you need to remember anything from first year/high school mathematics (i.e basis calculus, log laws, basic vector/matrix identities).

Some general identities which may be useful for this course:

Vector Calculus

If x is an arbitrary vector, and c is any constant (vector or scalar),

$$\frac{\partial(xc)}{\partial x} = c^T \qquad \frac{\partial(x^T cx)}{\partial x} = 2cx$$

The First Question

What is this problem, and how do we solve it?

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|_2^2$$

The First Question

What is this problem, and how do we solve it?

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|_2^2$$

Describe Ridge and LASSO regression and how they differ.

Linear Methods

Name this algorithm and what it represents,

$$\begin{aligned}\hat{p} &= \sigma(X\beta) \\ &= \frac{1}{1+e^{-X\beta}}\end{aligned}$$

Dual Perceptron

Recall the primal perceptron:

$converged \leftarrow 0$
while not $converged$ **do**
  $converged \leftarrow 1$
  for $x_i \in X, y_i \in y$ **do**
    if $y_i w \cdot x_i \leq 0$ **then**
      $w \leftarrow w + \eta y_i x_i$
      $converged \leftarrow 0$
    end if
  end for
end while

Dual Perceptron

Recall the primal perceptron:

$converged \leftarrow 0$
while not $converged$ **do**
  $converged \leftarrow 1$
  for $x_i \in X, y_i \in y$ **do**
    if $y_i w \cdot x_i \leq 0$ **then**
      $w \leftarrow w + \eta y_i x_i$
      $converged \leftarrow 0$
    end if
  end for
end while

- How did we derive the dual perceptron?

Dual Perceptron

Recall the primal perceptron:

$converged \leftarrow 0$
while not $converged$ **do**
  $converged \leftarrow 1$
  for $x_i \in X, y_i \in y$ **do**
    if $y_i w \cdot x_i \leq 0$ **then**
      $w \leftarrow w + \eta y_i x_i$
      $converged \leftarrow 0$
    end if
  end for
end while

- How did we derive the dual perceptron?
- What is the **Kernel trick**?

Dual Perceptron

Recall the primal perceptron:

$converged \leftarrow 0$
while not $converged$ **do**
  $converged \leftarrow 1$
  for $x_i \in X, y_i \in y$ **do**
    if $y_i w \cdot x_i \leq 0$ **then**
      $w \leftarrow w + \eta y_i x_i$
      $converged \leftarrow 0$
    end if
  end for
end while

- How did we derive the dual perceptron?
- What is the **Kernel trick**?
- What problem does the SVM solve?

Ensemble Methods

Describe the difference between bagging and boosting.

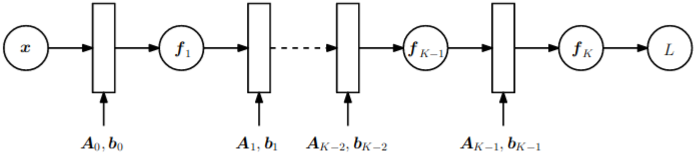
Ensemble Methods

Describe the difference between bagging and boosting.

Why does bagging reduce our model's variance?

Neural Learning

Given the following diagram, derive expressions for $\frac{\partial L}{\partial \theta_k}$ for $k = 0, \ldots, K$ where $\theta_k = \{A_k, b_k\}$



Section 3

Gradient Descent Question

Gradient Descent Question

Given $w = (w_0, w_1, w_2, w_3)^T$, $X^{(i)} = (1, x_1^{(i)}, x_2^{(i)}, x_3^{(i)})$ for a model:

$$\hat{y}^i = w_0 + w_1 x_1^{(i)} + w_2 x_2^{(i)} + w_3 x_3^{(i)} \hat{y}^i \qquad = w^T X^{(i)}$$

We define the mean-loss of our model as:

$$L_c(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} L_c(y^{(i)}, \hat{y}^{(i)}) = \frac{1}{n} \sum_{i=1}^{n} \left[\sqrt{\frac{1}{c^2} (y^{(i)} - \langle w^{(t)}, X^{(i)} \rangle)^2 + 1} - 1 \right]$$

Part A

Calculate $\frac{\partial L_c(y,\hat{y})}{\partial w_k}$, where $k = 0, \ldots, 4$.

Part B

Take $c = 2$, what are the GD updates to w for a learning rate η? What are the SGD updates?

Part B

Take $c = 2$, what are the GD updates to w for a learning rate η? What are the SGD updates?

$$w_k^{(t+1)} = w_k^{(t)} - \eta \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{X_k^{(i)}(y_i - \langle w^{(t)}, X^{(i)} \rangle)}{2\sqrt{(y_i - \langle w^{(t)}, X^{(i)} \rangle)^2 + 4}}$$

Part B

Take $c = 2$, what are the GD updates to w for a learning rate η? What are the SGD updates?

$$w_k^{(t+1)} = w_k^{(t)} - \eta \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{X_k^{(i)}(y_i - \langle w^{(t)}, X^{(i)} \rangle)}{2\sqrt{(y_i - \langle w^{(t)}, X^{(i)} \rangle)^2 + 4}}$$

For SGD,

$$w_k^{(t+1)} = w_k^{(t)} - \frac{X_k^{(i)}(y_i - \langle w^{(t)}, X^{(i)} \rangle)}{2\sqrt{(y_i - \langle w^{(t)}, X^{(i)} \rangle)^2 + 4}} \qquad \text{for a random } i \in [1, n]$$