COMP9417 - Week 10 Tutorial notes

Unsupervised Learning + Revision

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)

Revision (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),

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

The First Question

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

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|_2^2$$

Describe Ridge and LASSO regression and how they differ

Linear Methods

Name this algorithm and what it represents:

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

Dual Perceptron

Recall the primal perceptron:

```
converged ← 0
while not converged ← 1
    converged ← 1
    for x_i ∈ X, y_i ∈ y do
        if y_i w · x_i ≤ 0 then
            w ← w + η y_i x_i
            converged ← 0
        end if
    end for
end while
```

Pseudocode

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

Ensemble Methods

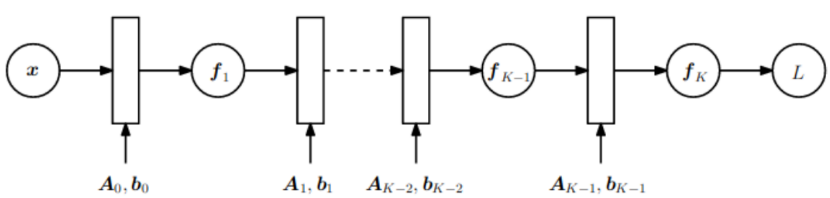
Describe the difference between bagging and boosting

Why does bagging reduce our model's variance?

Neural Learning

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

$\theta_k = \{A_k, b_k\}$



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:

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

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 n? What are the GD updates?

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

For SGD,

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