Practical Lecture 3 - Learning by Optimization

Machine Learning Course - 2nd Semester 2019/2020

Instituto Superior Tecnico, Universidade de Lisboa

1 Closed form learning

1) Consider the following training data:

$$\left\{\mathbf{x}^{(1)} = \begin{pmatrix}1\\1\end{pmatrix}, \mathbf{x}^{(2)} = \begin{pmatrix}2\\1\end{pmatrix}, \mathbf{x}^{(3)} = \begin{pmatrix}1\\3\end{pmatrix}, \mathbf{x}^{(4)} = \begin{pmatrix}3\\3\end{pmatrix}\right\}$$

$$\left\{t^{(1)} = 1.4, t^{(2)} = 0.5, t^{(3)} = 2, t^{(4)} = 2.5\right\}$$

a) Find the closed form solution for a linear regression that minimizes the sum of squared errors on the training data..

b) Predict the target value for $x_{query} = \begin{pmatrix}2\ 3\end{pmatrix}^T$.

c) Sketch the predicted hyperplane along which the linear regression predicts points will fall.

d) Compute the mean squared error produced by the the linear regression.

2) Consider the following training data:

$$\left\{\mathbf{x}^{(1)} = (-2.0), \mathbf{x}^{(2)} = (-1.0), \mathbf{x}^{(3)} = (0.0), \mathbf{x}^{(4)} = (2.0)\right\}$$

$$\left\{t^{(1)} = 2.0, t^{(2)} = 3.0, t^{(3)} = 1.0, t^{(4)} = -1.0\right\}$$

a) Find the closed form solution for a linear regression that minimizes the sum of squared errors on the training data..

b) Predict the target value for $x_{query} = (1)^T$.

c) Sketch the predicted hyperplane along which the linear regression predicts points will fall.

d) Compute the mean squared error produced by the the linear regression.

3) Consider the following training data:

$$\left\{\mathbf{x}^{(1)} = \begin{pmatrix}1\\1\end{pmatrix}, \mathbf{x}^{(2)} = \begin{pmatrix}2\\1\end{pmatrix}, \mathbf{x}^{(3)} = \begin{pmatrix}1\\3\end{pmatrix}, \mathbf{x}^{(4)} = \begin{pmatrix}3\\3\end{pmatrix}\right\}$$

$$\left\{t^{(1)} = 1, t^{(2)} = 1, t^{(3)} = 0, t^{(4)} = 0\right\}$$

a) Find the closed form solution for a linear regression that minimizes the sum of squared errors on the training data..

b) Use your linear regression to classify $x_{query} = \begin{pmatrix}2\ 2.5\end{pmatrix}^T$, assuming a threshold similarity of 0.5.

4) Consider the following training data:

$$\left\{\mathbf{x}^{(1)} = (-2.0), \mathbf{x}^{(2)} = (-1.0), \mathbf{x}^{(3)} = (0.0), \mathbf{x}^{(4)} = (2.0)\right\}$$

$$\left\{t^{(1)} = 1, t^{(2)} = 0, t^{(3)} = 0, t^{(4)} = 0\right\}$$

a) Find the closed form solution for a linear regression that minimizes the sum of squared errors on the training data..

b) Use your linear regression to classify $x_{query} = (-0.3)^T$, assuming a threshold similarity of 0.15.

5) Consider the following training data:

$$\mathbf{x}^{(1)} = \begin{pmatrix} -0.95 \\ 0.62 \end{pmatrix}, \mathbf{x}^{(2)} = \begin{pmatrix} 0.63 \\ 0.31 \end{pmatrix}, \mathbf{x}^{(3)} = \begin{pmatrix} -0.12 \\ -0.21 \end{pmatrix}, \mathbf{x}^{(4)} = \begin{pmatrix} -0.24 \\ -0.5 \end{pmatrix},$$
$$\mathbf{x}^{(5)} = \begin{pmatrix} 0.07 \\ -0.42 \end{pmatrix}, \mathbf{x}^{(6)} = \begin{pmatrix} 0.03 \\ 0.91 \end{pmatrix}, \mathbf{x}^{(7)} = \begin{pmatrix} 0.05 \\ 0.09 \end{pmatrix}, \mathbf{x}^{(8)} = \begin{pmatrix} -0.83 \\ 0.22 \end{pmatrix}$$

$$\left\{t^{(1)} = 0, t^{(2)} = 0, t^{(3)} = 1, t^{(4)} = 0, t^{(5)} = 1, t^{(6)} = 0, t^{(7)} = 1, t^{(8)} = 0\right\}$$

a) Plot the data points and try to choose a non-linear transformation to apply.

b) Adopt the non-linear transform you chose in a) and find the closed form solution.

c) Sketch the predicted surface along which the predictions will fall.

6) Consider the following training data:

$$\left\{\mathbf{x}^{(1)} = (3), \mathbf{x}^{(2)} = (4), \mathbf{x}^{(3)} = (6), \mathbf{x}^{(4)} = (10), \mathbf{x}^{(5)} = (12)\right\}$$

$$\left\{t^{(1)} = 1.5, t^{(2)} = 9.3, t^{(3)} = 23.4, t^{(4)} = 45.8, t^{(5)} = 60.1\right\}$$

a) Adopt a logarithmic feature transformation $\phi(x_1) = \log(x_1)$ and find the closed form solution for this non-linear regression that minimizes the sum of squared errors on the training data.

b) Repeat the exercise above for a quadratic feature transformation $\phi(x_1) = x_1^2$.

c) Plot both regressions.

d) Which is a better fit a) or b)?

2 Gradient descent learning

1) Consider the following training data:

$$\left\{\mathbf{x}^{(1)} = \begin{pmatrix}1\\1\end{pmatrix}, \mathbf{x}^{(2)} = \begin{pmatrix}2\\1\end{pmatrix}, \mathbf{x}^{(3)} = \begin{pmatrix}1\\3\end{pmatrix}, \mathbf{x}^{(4)} = \begin{pmatrix}3\\3\end{pmatrix}\right\}$$

$$\left\{t^{(1)} = 1, t^{(2)} = 1, t^{(3)} = 0, t^{(4)} = 0\right\}$$

In this exercise, we will work with a unit that computes the following function:

$$output\left(\mathbf{x}; \mathbf{w}\right) = \frac{1}{1 + \exp\left(-2\mathbf{w} \cdot \mathbf{x}\right)}$$

And we will use the half sum of squared errors as our error (loss) function:

$$E\left(\mathbf{w}\right) = \frac{1}{2}\sum_{k=1}^{N}\left(t^{(k)} - output\left(\mathbf{x}^{(k)}; \mathbf{w}\right)\right)^2$$

a) Determine the gradient descent learning rule for this unit.

b) Compute the first gradient descent update assuming an initialization of all ones .

c) Compute the first stochastic gradient descent update assuming an initialization of all ones.

2) Consider the following training data:

$$\left\{\mathbf{x}^{(1)} = \begin{pmatrix}1\\1\end{pmatrix}, \mathbf{x}^{(2)} = \begin{pmatrix}2\\1\end{pmatrix}, \mathbf{x}^{(3)} = \begin{pmatrix}1\\3\end{pmatrix}, \mathbf{x}^{(4)} = \begin{pmatrix}3\\3\end{pmatrix}\right\}$$

$$\left\{t^{(1)} = 1, t^{(2)} = 1, t^{(3)} = 0, t^{(4)} = 0\right\}$$

In this exercise, we will work with a unit that computes the following function:

$$output\left(\mathbf{x}; \mathbf{w}\right) = \frac{1}{1 + \exp\left(-\mathbf{w} \cdot \mathbf{x}\right)}$$

And we will use the cross-entropy loss function:

$$E\left(\mathbf{w}\right) = -\log\left(p\left(\mathbf{t} \mid \mathbf{w}\right)\right) = -\sum_{k=1}^{N}\left(t^{(k)}\log output^{(k)}\left(\mathbf{x}^{(k)}; \mathbf{w}\right) + \left(1 - t^{(k)}\right)\log\left(1 - output^{(k)}\left(\mathbf{x}^{(k)}; \mathbf{w}\right)\right)\right)$$

a) Determine the gradient descent learning rule for this unit.

b) Compute the first gradient descent update assuming an initialization of all ones .

c) Compute the first stochastic gradient descent update assuming an initialization of all ones.

3) Consider the following training data:

$$\left\{\mathbf{x}^{(1)} = \begin{pmatrix}1\\1\end{pmatrix}, \mathbf{x}^{(2)} = \begin{pmatrix}2\\1\end{pmatrix}, \mathbf{x}^{(3)} = \begin{pmatrix}1\\3\end{pmatrix}, \mathbf{x}^{(4)} = \begin{pmatrix}3\\3\end{pmatrix}\right\}$$

$$\left\{t^{(1)} = 1, t^{(2)} = 1, t^{(3)} = 0, t^{(4)} = 0\right\}$$

In this exercise, we will work with a unit that computes the following function:

$$output\left(\mathbf{x}; \mathbf{w}\right) = \exp\left(\left(\mathbf{w} \cdot \mathbf{x}\right)^2\right)$$

And we will use the half sum of squared errors as our error (loss) function:

$$E\left(\mathbf{w}\right) = \frac{1}{2}\sum_{k=1}^{N}\left(t^{(k)} - output\left(\mathbf{x}^{(k)}; \mathbf{w}\right)\right)$$

a) Determine the gradient descent learning rule for this unit.

b) Compute the stochastic gradient descent update for input $\mathbf{x} = \begin{pmatrix}1\\1\end{pmatrix}, t = 0$ initialized with $\mathbf{w} = \begin{pmatrix}0\\1\\0\end{pmatrix}$ and learning rate $\eta = 2$.

3 Thinking Questions

a) Until now we could only solve classification tasks where the two classes were separated by simple lines. Now we have seen that we can apply any feature transformations we want. Think about which kinds of problems we can solve now? Is it all a matter of finding the right transformation? Is it easy to choose the right transformation?

b) When using the linear regression for classification, think about how the threshold changes the sensitivity of the model. Is it more or less likely that the model will fail to recognize a class member as the threshold increases?

c) Think about the error functions we have seen. Do you think that one is clearly better than the other? What changes when one changes the error function?