ECS 171 Sample exercises in Machine Learning - UC Davis, 2017 Fall

Instructor: Ilias Tagkopoulos

December 3, 2017

1 Artificial Neural Networks

Problem 1. *Assume a single sample with* $\{x_1 = 0.5, x_2 = 0.8, y = 0.7\}$*. Build a model that predicts y from* x_1 *and* x_2 *using a feed-forward neural network with one hidden node (**Fig. 1**). Assume the g logistic function to be the activation function. Calculate weights updated using back-propagation and gradient descent. Assume that the initial weights are* $\{w_{10}^{(1)} = 0.5, w_{11}^{(1)} = 0.3, w_{12}^{(1)} = 0.2, w_{10}^{(2)} = 0.5, w_{11}^{(2)} = 0.8\}$ *and learning rate is 0.01.*

Solution. Forward propagation of the given sample $\{x_1 = 0.5, x_2 = 0.8, y = 0.7\}$ is as follows.

$$\begin{aligned} a_1^{(2)} &= g(w_{10}^{(1)} + x_1 w_{11}^{(1)} + x_2 w_{12}^{(1)}) \\ &= g(0.5 + 0.5 \cdot 0.3 + 0.8 \cdot 0.2) = 0.69 \\ a_1^{(3)} &= w_{10}^{(2)} + x_1 w_{11}^{(2)} + x_2 w_{12}^{(2)} \\ &= 0.5 + 0.69 \cdot 0.8 = 1.05 \end{aligned}$$

Errors are measured as follows:

$$\begin{aligned} \delta_1^{(3)} &= a_1^{(3)} - y = 0.35 \\ \delta_1^{(2)} &= w_1^{(2)T} \delta^{(3)} \cdot a_1^{(2)} (1 - a_1^{(2)}) \\ &= 0.8 \cdot 0.35 \cdot 0.69 \cdot 0.31 = 0.06 \end{aligned}$$

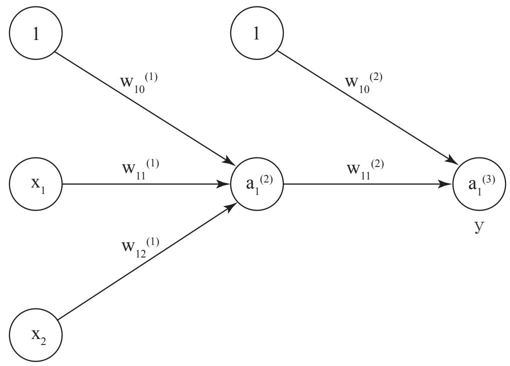

Figure 1: the architecture of the feedfoward neural network to predict y from x_1 and x_2 with one hidden node.

Weights are updated as follows:

$$\begin{aligned}
w_{10}^{(2)} &= w_{10}^{(2)} - \alpha \cdot a_0^{(2)} \cdot \delta_1^{(3)} \\
&= 0.5 - 0.01 \cdot 1 \cdot 0.35 = 0.496 \\
w_{11}^{(2)} &= w_{11}^{(2)} - \alpha \cdot a_1^{(2)} \cdot \delta_1^{(3)} \\
&= 0.8 - 0.01 \cdot 0.69 \cdot 0.35 = 0.79 \\
w_{10}^{(1)} &= w_{10}^{(1)} - \alpha \cdot a_0^{(1)} \cdot \delta_1^{(2)} \\
&= 0.5 - 0.01 \cdot 1 \cdot 0.06 = 0.499 \\
w_{11}^{(1)} &= w_{11}^{(1)} - \alpha \cdot a_1^{(1)} \cdot \delta_1^{(2)} \\
&= 0.3 - 0.01 \cdot 0.5 \cdot 0.06 = 0.299 \\
w_{12}^{(1)} &= w_{12}^{(1)} - \alpha \cdot a_2^{(1)} \cdot \delta_1^{(2)} \\
&= 0.2 - 0.01 \cdot 0.8 \cdot 0.06 = 0.199
\end{aligned}$$

2 Principal Component Analysis

Problem 2. *Find the covariance matrix of X where X is*

$$\begin{bmatrix} 1 & 2 \\ 2 & 4 \\ 3 & 3 \end{bmatrix}$$

Solution. First center the data by $X - \text{mean}(X)$.

$$\begin{bmatrix} 1-\mu_1 & 2-\mu_2 \\ 2-\mu_1 & 4-\mu_2 \\ 3-\mu_1 & 3-\mu_2 \end{bmatrix} = \begin{bmatrix} -1 & -1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Then measure the covariance of the centered data.

$$\begin{bmatrix} \frac{\Sigma x_1 x_1}{N-1} & \frac{\Sigma x_1 x_2}{N-1} \\ \frac{\Sigma x_2 x_1}{N-1} & \frac{\Sigma x_2 x_2}{N-1} \end{bmatrix} = \frac{X^T X}{N-1} = \frac{1}{2} \cdot \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

Problem 3. *Explain what PCA does and show how it is computed.*

Solution. It finds n orthogonal vectors which maximize the variance of samples when projected on these vectors. We call these vectors eigenvectors.

Problem 4. *What does PCA maximize?*

Solution.

$$\begin{aligned} &\textbf{maximize } \frac{1}{N-1} \sum_{i=1}^{m} (x^{(i)T} u)^2 \\ &\textbf{such that } ||u|| = 1 \end{aligned}$$

$$\begin{aligned} \frac{1}{N-1} \sum_{i=1}^{m} (x^{(i)T} u)^2 &= \frac{1}{N-1} \sum_{i=1}^{m} u^T x^{(i)} x^{(i)T} u \\ &= u^T \left(\frac{1}{N-1} \sum_{i=1}^{m} x^{(i)} x^{(i)T}\right) u \\ &= u^T \mathrm{cov}(x) u \\ &= u^T A u \end{aligned}$$

Find u that maximizes $u^T A u$ such that $||u|| = 1$. And it is done by finding the eigenvalues of A and plug in each eigenvalue to find respective eigenvector where

$$\begin{aligned} Au &= \lambda u \\ (A - \lambda I) u &= 0 \end{aligned}$$

λ can be computed from $|A - \lambda I| = 0$.

Problem 5. *Compute eigenvalues of A where A is*

$$\begin{bmatrix} 13 & 5 \\ 2 & 4 \end{bmatrix}$$

Solution.

$$\begin{aligned} |A - \lambda I| &= \begin{bmatrix} 13 - \lambda & 5 \\ 2 & 4 - \lambda \end{bmatrix} = (13 - \lambda)(4 - \lambda) - 5 \cdot 2 \\ &= \lambda^2 - 17\lambda + 52 - 10 = \lambda^2 - 17\lambda + 42 = 0 \end{aligned}$$

Thus, λ is either 3 or 14.

Problem 6. *Plug-in eigenvalues computed from problem 5 to find respective eigenvectors.*

Solution. For λ=3,

$$(A - \lambda I)u = \begin{bmatrix} 10 & 5 \\ 2 & 1 \end{bmatrix} = 0 = \begin{bmatrix} 10u_1 + 5u_2 = 0 \\ 2u_1 + u_2 = 0 \end{bmatrix}$$

Be aware that $||u|| = 1$ and thus $u_1 = -0.45$ and $u_2 = 0.9$. For λ=14,

$$(A - \lambda I)u = \begin{bmatrix} -1 & 5 \\ 2 & -10 \end{bmatrix} = 0 = \begin{bmatrix} -u_1 + 5u_2 = 0 \\ 2u_1 - 10u_2 = 0 \end{bmatrix}$$

Be aware that $||u|| = 1$ and thus $u_1 = 0.98$ and $u_2 = -0.2$.

3 Naive Bayes Classifier

Problem 7. *Given the following training data (**Table 1**), using a naive Bayes classifier, predict y of new sample $\{x_1 = S, x_2 = C, x_3 = H, x_4 = S\}$.*

x_1	x_2	x_3	x_4	y
S	H	H	W	N
S	H	H	S	N
O	H	H	W	Y
R	M	H	W	Y
R	C	N	W	Y
R	C	N	S	N
O	C	N	S	Y
S	M	H	W	N
S	C	N	W	Y
R	M	N	W	Y
S	M	N	S	Y
O	M	H	S	Y
O	H	N	W	Y
R	M	H	S	N

Table 1: Training data

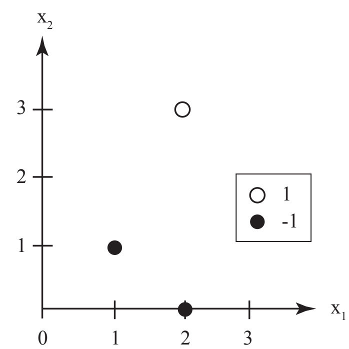

Figure 2: the training set for SVM.

Solution.

$$\begin{aligned}
c^* = & \underset{c=\{Y,N\}}{\textbf{argmax}} P(y=c|x_1=S,x_2=C,x_3=H,x_4=S) \\
= & \underset{c=\{Yes,No\}}{\textbf{argmax}} P(y=c|x_1=S,x_2=C,x_3=H,x_4=S) \\
= & \underset{c=\{Yes,No\}}{\textbf{argmax}} \frac{P(x_1=S,x_2=C,x_3=H,x_4=S|y=c)P(y=c)}{P(x_1=S,x_2=C,x_3=H,x_4=S)} \\
= & \underset{c=\{Yes,No\}}{\textbf{argmax}} P(x_1=S,x_2=C,x_3=H,x_4=S|y=c)P(y=c) \\
= & \underset{c=\{Yes,No\}}{\textbf{argmax}} P(x_1=S|y=c)P(x_2=C|y=c)P(x_3=H|y=c)P(x_4=S|y=c)P(y=c)
\end{aligned}$$

$$\begin{aligned}
P(x_1=S|y=Y)P(x_2=C|y=Y)P(x_3=H|y=Y)P(x_4=S|y=Y)P(y=Y) \\
= 0.22 \cdot 0.33 \cdot 0.33 \cdot 0.33 \cdot 0.64 = 0.0050 \\
P(x_1=S|y=N)P(x_2=C|y=N)P(x_3=H|y=N)P(x_4=S|y=N)P(y=N) \\
= 0.6 \cdot 0.2 \cdot 0.8 \cdot 0.6 \cdot 0.35 = 0.0201
\end{aligned}$$

Hence, the predicted y is N.

4 Support Vector Machines

Problem 8. *Build a SVM over the data set shown in* ***Fig. 2*** *($x^{(1)} = (1,1)$), $x^{(2)} = (2,3)$, $x^{(3)} = (2,0)$).*

Solution. We wolud like to find the line $w^T x + b = 0$ that maximizes the margin between the line $w^T x + b = -1$ and the line $w^T x + b = 1$ such that $y^{(i)}(w^T x^{(i)} + b) \geq 1$, for all i. In other words,

$$\begin{aligned}\textbf{minimize } & \frac{1}{2}||w||^2 \\ \textbf{such that } & y^{(i)}(w^T x^{(i)} + b) \geq 1, \text{ for all } i\end{aligned}$$

The Lagrangian form of the optimization problem is

$$\textbf{minimize } \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \alpha_i \{y^{(i)}(w^T x^{(i)} + b) - 1\}$$

Its derivative with respect to w and b is set to zero. Then we have

$$\begin{aligned} w &= \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \\ 0 &= \sum_{i=1}^{m} \alpha_i y^{(i)} \end{aligned}$$

There are two support vectors of $(1,1)$ and $(2,3)$ from visual inspection - just plot the points as in Figure 2 and you see they are the closest. In real problems the quadratic solver will pick the support vectors automatically based on the optimization procedure. Then we have

$$\begin{aligned} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} &= \begin{bmatrix} -\alpha_1 + 2\alpha_2 \\ -\alpha_1 + 3\alpha_2 \end{bmatrix} \\ 0 &= -\alpha_1 + \alpha_2 \end{aligned}$$

Then $w_1 = \alpha_2$ and $w_2 = 2\alpha_2$. By the constraint for support vectors, we have

$$\begin{aligned} \alpha_2 + 2\alpha_2 + b &= -1 \\ 2\alpha_2 + 6\alpha_2 + b &= 1 \end{aligned}$$

Therefore, $\alpha_2 = \frac{2}{5}$ and $b = -\frac{11}{5}$. So the optimal line is given by $w = (\frac{2}{5}, \frac{4}{5})$ and $b = -\frac{11}{5}$.