Introduction to Machine Learning **Fall Semester**

Recitation 10

Lecturer: Regev Schweiger *Scribe: Regev Schweiger*

10.1 Project

See project guidelines.

10.2 Linear Regression - Review

Least Squares

Our basic model is a linear function, i.e., $h_{\boldsymbol{\theta}}(\mathbf{x}) = \boldsymbol{\theta}^T\mathbf{x}$, where $\boldsymbol{\theta}$ are the parameters we like to learn for the linear function. The examples are $(\mathbf{x}_i, y_i)$. We can model this by a matrix

$$X = \begin{pmatrix} \cdots & \mathbf{x}_1 & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & \mathbf{x}_n & \cdots \end{pmatrix}$$

The labels are

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

We would like to find the parameters $\boldsymbol{\theta}$ that minimize the loss, namely

$$\begin{aligned} \min_{\boldsymbol{\theta}} \sum_{i=1}^{n} loss(h_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) &= \min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (h_{\boldsymbol{\theta}}(\mathbf{x}_i) - y_i)^2 && \text{assume square loss} \\ &= \min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (\boldsymbol{\theta}^T\mathbf{x}_i - y_i)^2 && \text{linear predictions} \\ &= \min_{\boldsymbol{\theta}} \|X\boldsymbol{\theta} - \mathbf{y}\|_2^2 \end{aligned}$$

Probabilistic Interpretation

The regression minimization criterion can be interpreted as a likelihood maximization criterion. We assume a model in which the relation between $\mathbf{x}$ and y can be described as

$$y = \boldsymbol{\theta}^T\mathbf{x} + \epsilon,$$

where $\boldsymbol{\theta}$ are unknown constants, and $\epsilon \sim N(0, \sigma^2)$, for some unknown value σ. Note that this can also be formulated as $y \sim N(\boldsymbol{\theta}^T \mathbf{x}, \sigma^2)$, and in fact, this formulation allows for generalisations such as logistic regression, as we will see.

Given a set of training data points $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, we can write the log likelihood of the model as

$$\log \mathcal{L}(\boldsymbol{\theta}, \sigma; \{(\mathbf{x}_i, y_i)\}) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \boldsymbol{\theta}^T \mathbf{x}_i)^2 - \frac{n}{2} \log(2\pi\sigma^2)$$

It is easy to see that the likelihood is maximized at the linear regression solution.

Gradient Descent Update

In the case of linear regression, the function that we are interested in minimizing is $f(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{\theta}^T \mathbf{x}_i)^2$ (we multiplied by 1/2 for simplicity). The gradient is $\nabla f(\boldsymbol{\theta}) = -X^t(\mathbf{y} - X\boldsymbol{\theta})$. Therefore, we start from a guess $\boldsymbol{\theta}_0$, and in each iteration we define:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha X^t(\mathbf{y} - X\boldsymbol{\theta}_k)$$

which is:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha \sum_{i=1}^{n} (y_i - h_{\boldsymbol{\theta}}(\mathbf{x}_i))\, \mathbf{x}_i$$

10.2.1 Logistic Regression

Assume we need to predict a Boolean outcome $y \in \{0, 1\}$. While this is a classical classification setting, we can still use regression. The problem is that the linear function of the regression can map values to be larger than 1 or below 0. we like to map any real number z to a value in $[0, 1]$. One such function is $g(z) = \frac{1}{1+e^{-z}}$.

We will now give a maximum likelihood interpretation and use it to learn θ.

$$h_{\boldsymbol{\theta}}(\mathbf{x}) = \Pr_{\boldsymbol{\theta}}[y = 1|\mathbf{x}] = g(\boldsymbol{\theta}^T \mathbf{x})$$

Our classifier will therefore return continuous results in the range $[0, 1]$, which are to be interpreted as the probability of $y = 1$. If we wish to predict a Boolean outcome, we may choose the value of y with a high probability (or, equivalently, round the predicted outcome).

We will learn $\boldsymbol{\theta}$ using ML model. The probabilities in the model will be

$$\Pr_{\boldsymbol{\theta}}[y|\mathbf{x}] = (h_{\boldsymbol{\theta}}(\mathbf{x}))^y (1 - h_{\boldsymbol{\theta}}(\mathbf{x}))^{1-y}$$

This is therefore another case where y is distributed according to a distribution dependent on $\boldsymbol{\theta}^T\mathbf{x}$. The likelihood function would be

$$\mathcal{L}(\theta; X, \mathbf{y}) = \Pr_{\boldsymbol{\theta}}[\mathbf{y}|X] = \prod_{i=1}^{m} (h_{\boldsymbol{\theta}}(\mathbf{x}_i))^{y_i} (1 - h_{\boldsymbol{\theta}}(\mathbf{x}_i))^{1-y_i}$$

The log likelihood is

$$\ell(\boldsymbol{\theta}; X, \mathbf{y}) = \sum_{i=1}^{m} y_i \log(h_{\boldsymbol{\theta}}(\mathbf{x}_i)) + (1 - y_i) \log(1 - h_{\boldsymbol{\theta}}(\mathbf{x}_i))$$

Gradient Ascent Update

In this case, we do not have a close form solution. We will therefore use gradient ascent to find the maximal solution.

Computing the derivative of g we have

$$g'(z) = \frac{e^{-z}}{(1 + e^{-z})^2} = \frac{1}{1 + e^{-z}} \cdot \frac{e^{-z}}{1 + e^{-z}} = g(z)(1 - g(z))$$

$$\begin{aligned}
\frac{d\ell}{d\boldsymbol{\theta}} &= \sum_{i=1}^{n} \frac{y_i}{g(\boldsymbol{\theta}^T\mathbf{x}_i)} g(\boldsymbol{\theta}^T\mathbf{x}_i)(1 - g(\boldsymbol{\theta}^T\mathbf{x}_i))\mathbf{x}_i - \frac{1 - y_i}{1 - g(\boldsymbol{\theta}^T\mathbf{x}_i)} g(\boldsymbol{\theta}^T\mathbf{x}_i)(1 - g(\boldsymbol{\theta}^T\mathbf{x}_i))\mathbf{x}_i \\
&= \sum_{i=1}^{n} \left(y_i(1 - g(\boldsymbol{\theta}^T\mathbf{x}_i)) - (1 - y_i) g(\boldsymbol{\theta}^T\mathbf{x}_i) \right) \mathbf{x}_i \\
&= \sum_{i=1}^{n} \left(y_i - g(\boldsymbol{\theta}^T\mathbf{x}_i) \right) \mathbf{x}_i
\end{aligned}$$

The update of $\boldsymbol{\theta}$ at the k^{th} iteration is:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha \sum_{i=1}^{n} \left(y_i - h_{\boldsymbol{\theta}}(\mathbf{x}_i) \right) \mathbf{x}_i$$

Note that this has the same form as the update for linear regression (however the updates are essentially different since the underlying $h_{\boldsymbol{\theta}}$ are different).