Machine Learning: Foundations **Fall semester, 2012**

Lecture 12: January 13, 2013

Lecturer: Yishay Mansour Scribe: Ofir Geri, Daniel Landau, Or Ozery, Rotem Zach[1]

12.1 Regression

12.1.1 General Problem

In previous lectures we focused on binary classification problems, where each sample had to be given one of two possible labels. Today we will discuss problems where the goal is to predict a real number, coming from a continuous range. For instance, we may wish to predict:

- How much will it rain tomorrow.
- How many years a person has to live.
- How much a stock will go up or down.

This problem of predicting real values is called *regression*.

We are given samples $S = \{(x_i, y_i)\}_{i=1}^m$ where x_i is sampled from distribution D and $y_i \in \mathbb{R}$. The values y_i are taken from a distribution that depends on x_i. Our goal is to build an hypothesis $h(x)$ that will be as close as possible to $f(x) = E[y|x]$. So we want to minimize $|h(x) - f(x)|$, but since we don't have access to $f(x)$, we'll minimize $|y - h(x)|$. The following claim will show that minimizing $|y - h(x)|$ also minimizes $|h(x) - f(x)|$.

Claim 12.1 $\underbrace{E_{x,y}[(y-h(x))^2]}_{\text{measured error}} = \underbrace{E_x[(f(x)-h(x))^2]}_{\text{prediction error}} + \underbrace{E_{x,y}[(y-f(x))^2]}_{\text{inherent uncertainty}}$

Proof:
Fix a value of x.

$$\begin{aligned} E_y[(y-h(x))^2] &= E_y[(y-f(x)+f(x)-h(x))^2] \\ &= E_y[(y-f(x))^2] + (f(x)-h(x))^2 + 2(f(x)-h(x))\underbrace{E_y[(y-f(x))]}_{0} \\ &= E_y[(y-f(x))^2] + (f(x)-h(x))^2 \end{aligned}$$

[1]Based on a scribes written by Gil Freundlich (June, 1996), Roi Yehoshua, Ophir Gvirtzer, Zohar Ganon (May, 2002) and Oana Sidi, Inbal Avraham and Vera Vsevolohzky (January, 2011).

Taking expectation over x,

$$E_{x,y}[(y-h(x))^2] \;=\; E_{x,y}[(y-f(x))^2] + E_x[(f(x)-h(x))^2]$$

■

In regression we are usually looking at the squared error, so given samples $S = \{(x_i, y_i)\}_{i=1}^m$, we would like to minimize $\frac{1}{m}\sum_{i=1}^m (y_i - h(x_i))^2$.

We take for example a class of linear hypotheses of the form $h(x) = wx$ where $x, w \in \mathbb{R}$. The error is $L = \frac{1}{m}\sum_{i=1}^m (y_i - x_i w)^2$. We'll take the derivative of L to find the best w:

$$\frac{\partial}{\partial w} L = \frac{1}{m}\sum_{i=1}^m -2x_i(y_i - x_i w) = 0 \longrightarrow w = \frac{\sum_{i=1}^m y_i x_i}{\sum_{i=1}^m x_i^2}$$

We can also solve this problem for hypotheses $h(x) = wx + b$ by adding a coordinate with value 1 to the x_i, having $h(x) = (w, b) \cdot (x, 1)$.

12.1.2 Linear Regression

A linear regressor is a mapping $x \mapsto w \cdot x$, where we assume that the instance space is a vector space (i.e., x is a vector) and the prediction is a linear combination of the instance vector x. The problem of learning a regression function with respect to a hypothesis class of linear predictors is called *linear regression*.

Formally, let $S = \{(x_1, y_1), (x_2, y_2)...(x_m, y_m)\}$ be a sequence of m training samples, where for each i we have $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$ (notice y_i takes continuous values). Consider the class of linear predictors:

$$\mathcal{H} \;=\; \{\; h_w(x) \;\;|\;\; w \in \mathbb{R}^n \;\},$$

where

$$h_w(x) = w \cdot x.$$

We shall find the best hypothesis with respect to the squared loss

$$L = \frac{1}{m}\sum_{i=1}^m (h_w x_i - y_i)^2 = \frac{1}{m}\sum_{i=1}^m (w \cdot x_i - y_i)^2.$$

We will choose to represent the problem as a vector w and a matrix X,

$$w \;=\; \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix}$$

$$X \;=\; \begin{pmatrix} - & x_1 & - \\ - & \vdots & - \\ - & x_m & - \end{pmatrix}$$

Where

$$\begin{aligned} x_i &\in \mathbb{R}^n \\ w &\in \mathbb{R}^n \\ y_i &\in \mathbb{R} \end{aligned}$$

Thus we can represent the loss as

$$L = (Xw - y)^T (Xw - y).$$

Rewriting the expression

$$L = (Xw)^T (Xw) - 2(Xw)y + y \cdot y.$$

We would like to minimize this loss, so we compute the gradien

$$\nabla L = 2X^T Xw - 2Xy = 0.$$

Therefore,

$$w = \left(X^T X\right)^{-1} X^T y.$$

This closed form solution is correct only when $X^T X$ is non-singular. If it is singular then one should use a pseudo-inverse (not discussed in this lecture).

12.1.3 Problems with Linear Regression

The least-squares solution we presented before might be highly non-stable - namely, a slight perturbation of the input causes a dramatic change of the output - leading again to overfitting.

Consider for example the case where $\mathcal{X} = \mathbb{R}^2$ and the training set contains two examples where the instances are $x_1 = (1, 0)$ and $x_2 = (1, \epsilon)$ and the targets are $y_1 = y_2 = 1$. The problem becomes,

$$P = \min_w \left[\frac{1}{2}(w_1 - 1)^2 + \frac{1}{2}(w_1 + \epsilon w_2 - 1)^2\right].$$

By vanishing the gradients we get the following system of equations:

$$\begin{cases} \frac{\partial P}{\partial w_1} = w_1 - 1 + w_1 + \epsilon w_2 - 1 = 0 \\ \frac{\partial P}{\partial w_2} = \epsilon(w_1 + \epsilon w_2 - 1) = 0 \end{cases}$$

With the solution:

$$\begin{cases} w_1 = 1 \\ w_2 = 0 \end{cases}$$

Now, let's repeat the above calculation with a slight change in the target: $y_1 = 1 + \epsilon$.

$$P = \min_w \left[\frac{1}{2}(w_1 - (1+\epsilon))^2 + \frac{1}{2}(w_1 + \epsilon w_2 - 1)^2 \right]$$

$$\begin{cases} \frac{\partial P}{\partial w_1} = w_1 - (1+\epsilon) + w_1 + \epsilon w_2 - 1 = 0 \\ \frac{\partial P}{\partial w_2} = \epsilon(w_1 + \epsilon w_2 - 1) = 0 \end{cases}$$

The solution is:

$$\begin{cases} w_1 = 1 + \epsilon \\ w_2 = -1 \end{cases}$$

That is, for some instances, a tiny change in the value of the targets makes a huge change in the solution (from $w_2 = 0$ to $w_2 = -1$) of the least squares estimator.

12.2 Regularization

A problem suffering from such instability is also called an ill-posed problem. A common solution is to add regularization. Regularization means adding a penalty for big weights. We shall see two kinds of regularization.

1. **Ridge Regression**

   We regularize by adding $\|w\|_2^2$ to the optimization problem, namely, to define the estimator as

   $$\arg\min_{w \in \mathbf{R}^n} \left(\frac{\lambda}{2}\|w\|_2^2 + \hat{SQ}_w \right), \tag{12.1}$$

   where λ is the regularization parameter and $\hat{SQ}_w = \sum_{i=1}^m \frac{1}{2}(w \cdot x_i - y_i)^2$ is the square loss.

   After solving Eq.(12.1) we obtain the following closed form solution for w:

   $$\begin{aligned} A &= X^T X + \lambda I \\ w &= A^{-1} X^T y \end{aligned}$$

   Since $X^T X$ is positive semi-definite, the matrix A has all its eigenvalues bounded from below by λ. Thus, all the eigenvalues of A^{-1} are bounded from above by $1/\lambda$, which guarantees a stable solution.

2. **Lasso Regression**

   Another form of regularization uses the l_1 norm. The resulting estimator is called Lasso:

$$\arg\min_{w\in\mathbf{R}^n}\left(\lambda\|w\|_1+\hat{SQ}_w\right), \tag{12.2}$$

   where again λ is the regularization parameter and $\hat{SQ}_w=\sum_{i=1}^m\frac{1}{2}(w\cdot x_i-y_i)^2$ is the square loss.

   While there is no closed form solution for the Lasso problem, it can still be solved efficiently by an of-the-shelf convex optimization method. In particular, we can apply the stochastic sub-gradient method for the Lasso problem. The advantage of using Lasso regression is that the resulting weights vector is many times sparse.

12.2.1 Generalization Error Bound for Ridge Regression

Note: The definition of ridge regression here has slight differences from the above definition. Consider the following optimization problem:

$$\min\sum\Psi_i^2$$

$$\text{s.t. } \Psi_i=w\cdot x_i-y_i\text{and } \|w\|_2^2\le\Lambda^2.$$

What is the solution's generalization ability? Let's assume that $\|x\|\le R$, $\|w\|\le\Lambda$. First we'll state the following Lemma (proof omitted):

Lemma 12.2 *[Ledoux-Talagrand] For $i=1,...,n$, let $\Phi_i:\mathbf{R}\to\mathbf{R}$ be an M-Lipschitz function with parameter M, i.e. $|\Phi_i(a)-\Phi_i(b)|\le M|a-b|,\forall a,b\in\mathbf{R}$. Then*

$$\hat{R}_S(\Phi\circ H)=\mathbf{E}_\sigma\left[\sup_{h\in H}\left|\frac{1}{m}\sum_{i=1}^m\sigma_i\Phi_i(h(x_i))\right|\right]\le M\mathbf{E}_\sigma\left[\sup_{h\in H}\left|\frac{1}{m}\sum_{i=1}^m\sigma_i h(x_i)\right|\right]=M\hat{R}_S(H)$$

where $\hat{R}_S$ is the Rademacher complexity.

We are interested in bounding the Rademacher complexity of the SQ class so we can obtain generalization error bound.
Let's assume that $\Psi_i=w\cdot x_i-y_i$ are bounded from above by M. Denote by $\Phi(\Psi)=\Psi^2$ then the function Φ is 2M-Lipschitz, and by applying Lemma 12.2 we get:

$$\hat{R}_S(SQ)\le 2M\hat{R}_S(Linear)\le 2M\frac{R\Lambda}{\sqrt{m}}$$

Where the last inequality was proven in a homework assignment. The following bound holds with high probability $\geq 1 - \delta$.

$$\mathbf{E}[SQ] \leq \hat{SQ}(S) + 2\hat{R}_S(SQ) + O\left(\sqrt{\frac{\ln \frac{1}{\delta}}{m}}\right)$$

12.2.2 Regression from a Bayesian Perspective

Assume the following process:

- Choose w from the distribution $N(0, \sigma)$
- Choose $Noise$ from the distribution $N(0, \sigma_0)$
- Given x , set $y = w \cdot x + Noise$

Bayes' Theorem states that:

$$Pr[w|(x_1, y_1), ..., (x_n, y_n)] = \frac{Pr[(x_1, y_1), ..., (x_n, y_n)|w]Pr[w]}{Pr[(x_1, y_1), ..., (x_n, y_n)]}$$

Within the Bayesian approach we want to get the MAP and ML.

ML:

$$\begin{aligned} w_{ML} &= \arg\max_w Pr[(x_1, y_1), ..., (x_n, y_n)|w] \\ &= \arg\max_w \prod_{i=1}^{n} Pr[(x_i, y_i)|w] \\ &= \arg\max_w e^{\sum_{i=1}^{n} -\frac{1}{2\sigma^2}(w \cdot x_i - y_i)^2} \\ &= \arg\min_w \sum_{i=1}^{n} \frac{1}{2\sigma^2}(w \cdot x_i - y_i)^2 \\ &= \arg\min_w \sum_{i=1}^{n} (w \cdot x_i - y_i)^2 \end{aligned}$$

We obtain the standard linear regression problem.

MAP:

$$\begin{aligned} w_{MAP} &= \arg\max_w Pr[(x_1, y_1), ..., (x_n, y_n)|w]Pr[w] \\ &= \arg\max_w e^{\sum_{i=1}^{n} -\frac{1}{2\sigma^2}(w \cdot x_i - y_i)^2} e^{\frac{\|w\|^2}{{\sigma_0}^2}} \end{aligned}$$

$$
\begin{aligned}
&= \arg\min_w \sum_{i=1}^{n} \frac{1}{2\sigma^2}(w \cdot x_i - y_i)^2 + \frac{\|w\|^2}{\sigma_{0^2}} \\
&= \arg\min_w \sum_{i=1}^{n} (w \cdot x_i - y_i)^2 + \lambda \|w\|^2
\end{aligned}
$$

where $\lambda = \frac{\sigma^2}{\sigma_0^2}$.

We obtain the standard ridge regression problem, and λ, the regularization parameter is the ratio of the two variances.

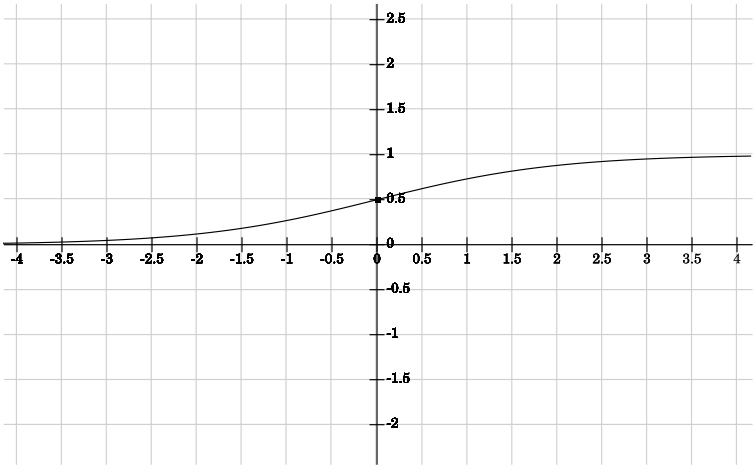
12.2.3 Logistic Regression

Recall that in binary classification problems, we wish to assign each instance a label from $Y = \{+1, -1\}$ or $Y = \{0, 1\}$. In such occasions, we may also be interested in the value of $Pr[Y = 1|X]$. If we use $Y = \{0, 1\}$, this value is the expectancy of the label y. We wish to estimate $Pr[y = 1|x]$ as $w \cdot x$, when we choose w as $argmin_w \sum_{i=1}^{n} (w \cdot x_i - y_i)^2$. We get that: $E[(w \cdot x_i - y_i)^2] = E[(w \cdot x - Pr[y = 1|x])^2]$. However, in that case, the value of $w \cdot x$ may not be in [0,1].

Instead, we can choose to use the following estimation:

$$Pr[Y = 1|X] = \sigma(w \cdot x),$$

where $\sigma(z) = \frac{1}{1+e^{-z}}$ (a plot of this function follows).



The function image is (0,1), and for every z, $\sigma(z) + \sigma(-z) = 1$. That is useful, as when we use the labels $Y = \{+1, -1\}$, we get that

$$Pr[Y = -1|X] = \sigma(-w \cdot x) = 1 - \sigma(w \cdot x),$$

or generally, for $y \in \{+1, -1\}$,

$$Pr[Y = y|x] = \sigma(yw \cdot x).$$

We shall now find the optimal w. We wish to minimize

$$\sum_{i=1}^{m} (\sigma(w \cdot x_i) - y_i)^2$$

This function is not convex, so it might have a local minimum. We will look at the following function instead:

$$\begin{aligned}
\hat{w} &= \arg\max_w \prod_{i=1}^{m} Pr[y_i|x_i, w] \\
&= \arg\max_w \sum_{i=1}^{m} log Pr[y_i|x_i, w] \\
&= \arg\min_w \sum_{i=1}^{m} -log\sigma(y_i w \cdot x_i) \\
&= \arg\min_w \sum_{i=1}^{m} log(1 + e^{y_i w \cdot x_i})
\end{aligned}$$

This function is convex, so we can solve$w$ using convex optimization.
The probabilities are

$$\begin{aligned}
Pr[Y = 1|X] &= \sigma(w \cdot x) \\
Pr[Y = -1|X] &= \sigma(-w \cdot x)
\end{aligned}$$

We may also turn this into a classifier: An instance is positive if $\sigma(w \cdot x) \geq \frac{1}{2}$, that is, $w \cdot x \geq 0$. This is a linear classifier, that is yet another classifer that fits a hyperplane.