STAT 391 Lecture 8
May 2023

**Linear and logistic regression**



mmp@stat.washington.edu

The task of **Prediction** is concerned with the relationship between two random variables, the **predictor** $X \in S_X$, and the **response** or **target** $Y \in S_Y$. The task is to predict the value of $Y$ that "best" corresponds to a given $X$. Therefore, statistically speaking, we are interested in (estimating) the conditional distribution $P_{Y|X}$.

When the outcome space of $Y$, $S_Y$ is a finite discrete set, prediction is called **classification**; when $S_Y \subset (-\infty, \infty)$, it is called **regression**.

## 1 Linear regression with a single predictor

Let $S_X = (-\infty, \infty)$. We assume a *linear model*, i.e.

$$y = \beta_0 + \beta_1 x + \epsilon, \tag{1}$$

where $\beta_{0,1} \in \mathbb{R}$ are called model **parameters** or **regression coefficients**, and $\epsilon$ is called **noise**. The noise $\epsilon$ makes the dependence of $Y$ on $X$ random, without it it will be deterministic. We assume that

$$\epsilon \sim Normal(0, \sigma^2), \tag{2}$$

and moreover, that for each value pair $(x, y)$ observed, the noise is independent of other observations.

We want to estimate the unknown parameters $\beta_0, \beta_1, \sigma^2$ by ML, from a data set $\mathcal{D} = \{(x^1, y^1), \ldots (x^n, y^n)\}$ sampled i.i.d. from an unknown distribution $P_{Y|X}$. Hence, we are not interested in the distribution of the $x_{1:n}$ variables, but only in the probabilistic depence of $Y$ on $X$. Note that our *model* for this distribution, based on (**??**) and (2) is

$$P_{Y|X} \;=\; Normal(\underbrace{\beta_0 + \beta_1 X}_{\mu(X)}, \sigma^2). \tag{3}$$

The likelihood function is defined as

$$L(\beta_{0,1}, \sigma^2) \;=\; P[y^{1:n}|x^{1:n}, \beta_{0,1}, \sigma^2] \;=\; \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y^i-\mu(x^i))^2}{\sigma^2}} \;=\; \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y^i-\mu(x^i))^2}, \tag{4}$$

and the log-likelihood is

$$l(\beta_{0,1}, \sigma^2) \;=\; \ln P[y_{1:n}|x_{1:n}, \beta_{0,1}, \sigma^2] \;=\; -n\ln\sigma - n\ln(\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y^i - \beta_0 - \beta_1 x^i)^2. \tag{5}$$

This reminds of the ML estimation of a normal distribution, so we proceed to first estimate the parameters $\beta_0, \beta_1$ of the mean.

$$\frac{\partial l}{\partial \beta_0} \;=\; \sum_{i=1}^{n}(y^i - \beta_0 - \beta_1 x^i) \tag{6}$$

$$\frac{\partial l}{\partial \beta_1} \;=\; \sum_{i=1}^{n} x^i(y^i - \beta_0 - \beta_1 x^i) \tag{7}$$

By setting the above partial derivatives to 0, we get the linear system

$$\sum_{i=1}^{n} y^i \;=\; n\beta_0 - \beta_1 \sum_{i=1}^{n} x^i \tag{8}$$

$$\sum_{i=1}^{n} x^i y^i \;=\; n\beta_0 \sum_{i=1}^{n} x^i - \beta_1 \sum_{i=1}^{n}(x^i)^2, \tag{9}$$

with solution

$$\beta_1^{ML} \;=\; \frac{n\sum_{i=1}^{n} x^i y^i - (\sum_{i=1}^{n} x^i)(\sum_{i=1}^{n} y^i)}{n\sum_{i=1}^{n}(x^i)^2 - (\sum_{i=1}^{n} x^i)^2} \tag{10}$$

$$\beta_0^{ML} \;=\; \frac{1}{n}\sum_{i=1}^{n} y^i - \beta_1^{ML}\frac{1}{n}\sum_{i=1}^{n} x^i \;=\; \bar{y} - \beta_1^{ML}\bar{x}. \tag{11}$$

## 2 Linear regression with multiple predictors

Let $X$ now be a vector variable, $X = (X_1, \ldots X_m) \in \mathbb{R}^m$. We assume $Y$ is a linear combination of all the $m$ predictors, i.e.

$$y \;=\; \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots \beta_m x_m + \epsilon. \tag{12}$$

This expression can be written more compactly in vector form, if we augment the vector $X$ with an additional component $X_0 \equiv 1$, i.e. $X \leftarrow (1, X_1, \ldots X_m) \in \mathbb{R}^{m+1}$. With this artifice, $\beta_0$ can be treated similarly with the other regression coefficients, which are all collected in the vector $\beta = [\beta_0 \, \beta_1 \, \ldots \beta_m]^T \in \mathbb{R}^{m+1}$. Now (12) becomes

$$y = \underbrace{\beta^T x}_{\mu(x)} + \epsilon. \tag{13}$$

Since the distribution of $\epsilon$ is given by (**??**), as before, the likelihood and log-likelihood are the same as in (4), respectively (5) with the only difference in the expression of $\mu(X)$.

$$l(\beta, \sigma^2) = \ln P[y^{1:n} | x^{1:n}, \beta, \sigma^2] = -n \ln \sigma - n \ln(\sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y^i - \beta^T x^i)^2. \tag{14}$$

If we ignore the first terms, which do not depend on $\beta$, we see that the parameters $\beta$ that maximize the (log-)likelihood are the ones that minimize the sum of squared **residuals** $y_i - \mu(x_i)$, hence this optimization is called a **least squares** problem.

We again take partial derivatives and equate them with 0. Remember that the partial derivative w.r.t. a vector variable $\beta$ is a vector called the *gradient*, and that this can be written as

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} (y^i - \beta^T x^i) x^i_j, \text{for all} j. \tag{15}$$

We can make this expression more compact if we construct the matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ with the $x^{1:n}$ as rows, and the column vector $\mathbf{y} = [y^1 \ldots y^n]^T$.

$$\frac{\partial l}{\partial \beta} = \mathbf{X}^T \mathbf{y} - \mathbf{X}^T \mathbf{X} \beta. \tag{16}$$

Setting the gradient to 0, we obtain the linear system $\mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y}$. If $n \geq m$, and the matrix $\mathbf{X}^T \mathbf{X}$ is non-singular, the solution is

$$\beta^{ML} = \underbrace{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T}_{\mathbf{X}^\dagger} \mathbf{y}. \tag{17}$$

The matrix $\mathbf{X}^\dagger$ is called the **pseudoinverse** of $\mathbf{X}$.

Once $\beta^{ML}$ is obtained, we can also estimate the residuals

$$\epsilon^i = y^i - (\beta^{ML})^T x^i. \tag{18}$$

# 3 Statistical properties of the $\beta^{ML}$ estimator

The expectation of $\beta^{ML}$ is computed w.r.t. the noise distribution, assuming that the data is generated by the model (13) (or (3)) with a true parameter vector $\beta$ and a true noise variance $\sigma^2$.

$$E[\beta^{ML}] \;=\; E[\mathbf{X}^\dagger \mathbf{y}] \;=\; E[\mathbf{X}^\dagger(\mathbf{X}\beta + \epsilon)] \;=\; \underbrace{\mathbf{X}^\dagger\mathbf{X}}_{I_m}\beta + \mathbf{X}^\dagger \underbrace{E[\epsilon]}_{0} \;=\; \beta. \tag{19}$$

The first equality is obtained by plugging in $\beta^{ML} = \mathbf{X}^\dagger\mathbf{y}$, and the second by replacing $\mathbf{y}$ with its values from the true model. We see from (19) that the ML estimate $\beta^{ML}$ is **unbiased**.

We can also calculate the covariance of $\beta^{ML}$. Note that $\beta^{ML} - \beta = \mathbf{X}^\dagger\epsilon$. Hence,

$$\begin{aligned} Cov(\beta^{ML}) &= E[(\beta^{ML} - \beta)(\beta^{ML} - \beta)^T] \;=\; E[(\mathbf{X}^\dagger\epsilon)(\mathbf{X}^\dagger\epsilon)^T] && (20)\\ &= E[\mathbf{X}^\dagger\epsilon\epsilon^T(\mathbf{X}^\dagger)^T] \;=\; \mathbf{X}^\dagger E[\epsilon\epsilon^T](\mathbf{X}^\dagger)^T \;=\; \mathbf{X}^\dagger\sigma^2 I_n(\mathbf{X}^\dagger)^T && (21)\\ &= \sigma^2\mathbf{X}^\dagger(\mathbf{X}^\dagger)^T\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} && (22)\\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}. && (23) \end{aligned}$$

Above, we use the fact that $\mathbf{X}^T\mathbf{X}$ is a symmetric matrix, and so is its inverse. The covariance of $\beta^{ML}$ is proportional to the noise covariance.

# 4 Estimating $\sigma^2$

A naive way to estimate $\sigma^2$ is to average the squared residuals $(\sigma^2)^{naive} = \frac{1}{n}\sum_{i=1}^n (y^i - (\beta^{ML})^T x^i)^2$. We can also use the ML method, by taking the derivative of $l(\beta, \sigma^2)$ w.r.t. $\sigma^2$ (this is similar to ML estimation of $\sigma^2$ in a normal distribution).

$$\frac{\partial l}{\partial \sigma^2} \;=\; -n\frac{1}{\sigma^4} - \frac{1}{2\sigma^2}\sum_{i=1}^n (y^i - (\beta^{ML})^T x^i)^2 \;=\; 0. \tag{24}$$

If we solve this equation, we obtain

$$(\sigma^2)^{ML} \;=\; \frac{1}{n}\sum_{i=1}^n (y^i - (\beta^{ML})^T x^i)^2 \tag{25}$$

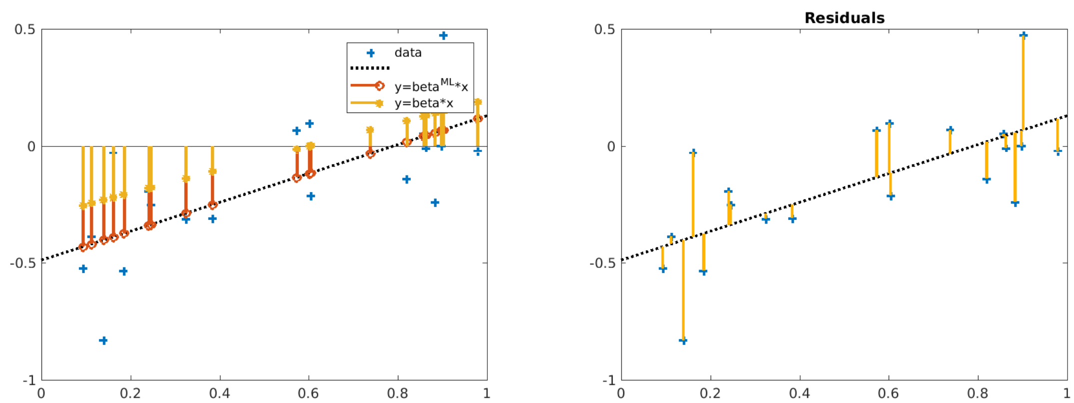


Figure 1: Left: Linear regression for $n = 20$ data points. The dotted line and circles are on the estimated regression line, while the yellow stars are on the true regression line, i.e. are the true $E[Y|X = x^i]$. Right: residuals $y^i - \beta_0^{ML} - \beta_1^{ML} x^i$.

which is identical to the "naive" estimator! However, just like in the case of the normal distribution, this estimator of $\sigma^2$ is also biased. By following the same procedure as in Chapter 12, we obtain

$$E[(\sigma^2)^{ML}] \;=\; \frac{n-m}{n}\sigma^2. \tag{26}$$

Therefore, unless $n \gg m$, the unbiased estimator

$$\hat{\sigma}^2 \;=\; \frac{1}{n-m}\sum_{i=1}^{n}(y^i - (\beta^{ML})^T x^i)^2 \;=\; \frac{n}{n-m}(\sigma^2)^{ML} \tag{27}$$

is recommended. (Note that here,$m$ is the number of total parameters estimated, , i.e., the dimension of $\beta$ with $\beta_0$ included.)

# 5 Prediction with the estimated model

Given a new $x$ value, the ML model for $P_{Y|X}(y|x)$ is $Normal(x^T\beta, (\sigma^2)^{ML})$, where we recall that $x^T\beta = \beta_0 + \beta_1 x_1 + \ldots \beta_m x_m$. This is the **predictive distribution** for $y$ given $x$.

If we want to predict a single number, given that the distribution is Gaussian, the "best" single number to predict is the mean $\mu(x) = x^T\beta$. [Exercise: in which ways is $\mu(x)$ "best"?] [Exercise: is $\mu(x)$ also "best" if we use the unbiased model $N(x\beta, \hat{\sigma}^2)$?]

# 6 Logistic Regression

When the outputs $y$ are binary variable, i.e. $y \in \{0, 1\}$, fitting them with a linear model is not appropriate. **Exercise:** *Why?* **Logistic regression** proposes that, for each $x$, the model for $P(Y|X)$ be a Bernoulli distribution, with $p(x) \stackrel{def}{=} Pr[Y = 1|X = x]$ given implicitly by the relation below.

Let $\beta$ be the vector of parameters as described above (with or without a $\beta_0$ included). The let $f(x) = \beta^T x$ model the **log odds** of class 1

$$f(X) \;=\; \frac{P(Y=1|X)}{P(Y=0|X)} \;=\; \beta^T X. \tag{28}$$

Then under this linear model, $p(x)$ is

$$\frac{p(x)}{1-p(x)} \;=\; e^{f(x)} \tag{29}$$

$$Pr[Y=1|X=x] \;=\; p(x) \;=\; \frac{e^f}{1+e^f} \;=\; \frac{e^{\beta^T x}}{1+e^{\beta^T x}} \;=\; \frac{1}{1+e^{-\beta^T x}} \tag{30}$$

$$Pr[Y=0|X=x] \;=\; 1-p(x) \;=\; \frac{e^{-\beta^T x}}{1+e^{-\beta^T x}} \;=\; \frac{1}{1+e^{\beta^T x}} \tag{31}$$

An alternative "symmetric" expression for $p, 1-p$ is

$$p \;=\; \frac{e^{f/2}}{e^{f/2}+e^{-f/2}}, \quad 1-p \;=\; \frac{e^{-f/2}}{e^{f/2}+e^{-f/2}}. \tag{32}$$

In the expression (**??**) one recognizes the *logistic CDF*. Expressions (30) and (31) can be written simultaneously as

$$Pr[Y|X=x] \;=\; \frac{e^{Y\beta^T x}}{1+e^{\beta^T x}} \tag{33}$$

One major application of logistic regression is in *classification*.

# 7 Estimating the parameters by Max Likelihood

The log-likelihood $l(\beta)$ is

$$\begin{aligned} l(\beta) &= \ln Pr[y^{1:n}|x^{1:n}, \beta] && (34) \\ &= \sum_{i=1}^{n} \ln \frac{e^{y^i \beta^T x^i}}{1 + e^{\beta^T x^i}} && (35) \\ &= \sum_{i=1}^{n} \left[ y^i \beta^T x^i - \ln(1 + e^{\beta^T x^i}) \right] && (36) \end{aligned}$$

There is no analytic formula for the maximum of this expression. Therefore, the Maximum Likelihood parameters $\beta^{ML}$ will be found numerically, by gradient ascent.

We first calculate the gradient of the log-likelihood.

$$\begin{aligned} \frac{\partial l}{\partial \beta_j} &= \sum_{i=1}^{n} \left[ y^i x_j^i - \frac{e^{\beta^T x^i}}{1 + e^{\beta^T x^i}} x_j^i \right] && (37) \\ &= \sum_{i=1}^{n} \left[ y^i - p(x^i) \right] x_j^i && (38) \end{aligned}$$

This expression can be written compactly for all $j = 0 : p$ as

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{n} \underbrace{\left[ y^i - p(x^i) \right]}_{c_i \in \mathbb{R}} x^i. \tag{39}$$

Recall that in gradient ascent, at every step,

$$\beta \leftarrow \beta + \eta \frac{\partial l}{\partial \beta}, \tag{40}$$

with $\eta > 0$ the *step size*. The expression of the gradient in (39) shows that the change in $\beta$, at each step, is a sum of vectors, each of them being a scaled version of a data point $x^i$. Hence, if the initial value of $\beta$ is zero, the parameter vector $\beta$ is at any time a *a linear combination* of the inputs $x^i$.

Next, we note that

$$c_i = y^i - p(x^i) = (-1)^{1-y^i} \left( 1 - Pr[y^i|x^i, \beta] \right); \tag{41}$$

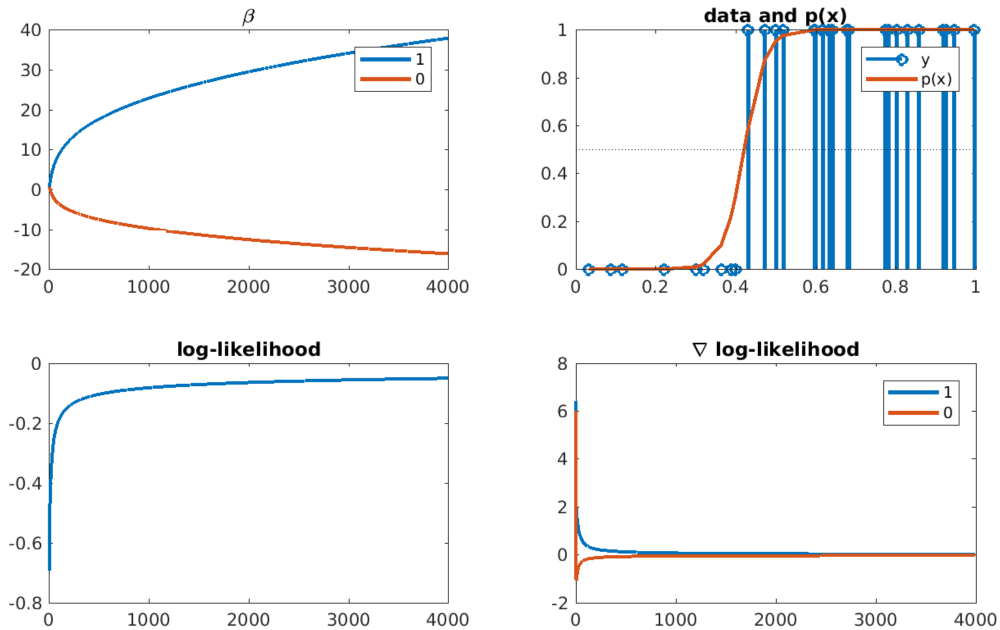


Figure 2: Logistic regression estimation by gradient ascent for $n = 30$ data points, $4,000$ iterations. Top left: $\beta_0, \beta_1$ trajectories; bottom left: log-likelihood; bottom right: derivatives $\frac{\partial l}{\partial \beta_0}, \frac{\partial l}{\partial \beta_1}$; to right data $(x^{1:30}, y^{1:30})$ and probability of $Y = 1$, $p(X)$, according to estimated model.

in other words, $|c_i|$ is the difference between the *ideal* prediction probability 1 and the model's probability of the observed $y^i$. Hence, for the data points $i$ for which the model predicts the outputs well, $|c_i|$ is close to 0. This leave the data points when the model is not accurate, to dominate in the gradient expression. We can also see that $c_i > 0$ when $y^i = 1$, and $c_i < 0$ when $y^i = 0$. In other words, each gradient step moves $\beta$ in the general direction of the $y^i = 1$ points (also called **positive examples**) and away from the $y^i = 0$ points (the **negative examples**).

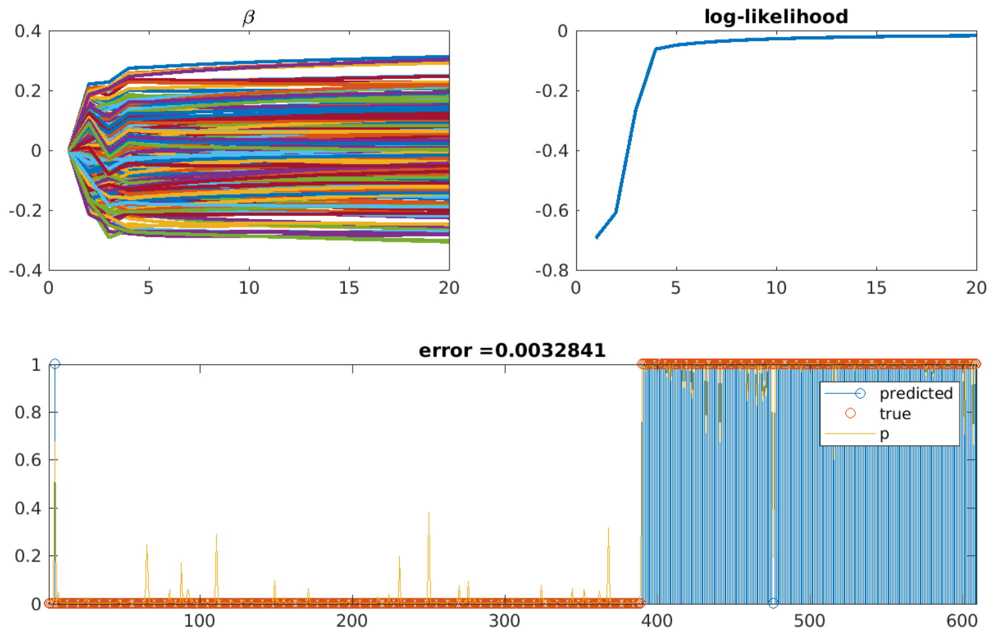


Figure 3: Logistic regression estimation by gradient ascent for $n = 609$ handwritted 0's and 2's in $d = 256$ dimensions, 20 iterations. Top left: trajectories for $\beta_{0:256}$; top right: log-likelihood; bottom: data $(1 : 609, y^{1:609})$ and probability of $Y = 1$, $p(X)$, according to estimated model.

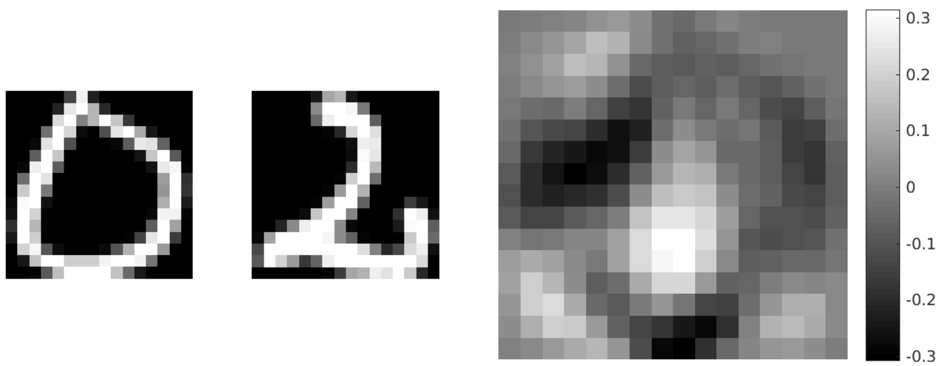


Figure 4: Two examples of handwritten digits from the data set; the parameters $\beta_{1:256}$ corresponding to each of the 256 pixels in a digit image.