CS 760: Machine Learning Fall 2020

Topic 6: Logistic Regression

Instructor: Daniel L. Pimentel-Alarcón

GO GREEN. AVOID PRINTING, OR PRINT 2-SIDED OR MULTIPAGE.

6.1 Introduction

Arguably the simplest classification task that we can teach a computer is to distinguish between two classes. For example:

1. Does this image contain a dog or a cat?
2. Is this person healthy or diabetic?
3. Would this individual survive a disaster?

Logistic regression is one of the most elemental yet powerful techniques for this purpose. The main idea is to compute the *likelihood* that a sample (e.g., a person) belongs to each class, based on its information and the information of previous (training) samples, and then choose the most likely class.

6.2 Setup

Suppose you want to determine whether you would have survived the Titanic sinking, based on certain information (features) about you, like age, gender, height, weight, etc. Let $\mathbf{x} \in \mathbb{R}^{\mathrm{D}}$ denote the feature vector containing this information, which may look like this:

$$\mathbf{x} = \begin{bmatrix} age \\ gender \\ height \\ weight \\ \vdots \end{bmatrix}. \tag{6.1}$$

Here D denotes the number of features. Similarly, let y be the random variable indicating whether you survive ($y = 1$) or not ($y = 0$). Hence we can rephrase our goal as determining whether $y = 0$ or $y = 1$ based on $\mathbf{x}$. Mathematically, we want to find a function f such that

$$y \;=\; f(\mathbf{x}).$$

Perhaps the most natural way to achieve this is to let f be of the form:

$$f(\mathbf{x}) \;=\; \begin{cases} 1 & \text{if } \mathbb{P}(y=1|\mathbf{x}) > \mathbb{P}(y=0|\mathbf{x}), \\ 0 & \text{otherwise.} \end{cases} \tag{6.2}$$

In words, (6.2) is simply saying: decide $y = 1$ if the probability of y being 1 (based on $\mathbf{x}$) is larger than the probability of y being 0, and decide $y = 0$ otherwise. We can rewrite (6.2) as follows:

$$f(\mathbf{x}) \;=\; \begin{cases} 1 & \text{if } \frac{\mathbb{P}(y=1|\mathbf{x})}{\mathbb{P}(y=0|\mathbf{x})} > 1, \\ 0 & \text{otherwise.} \end{cases}$$

The term $\frac{\mathbb{P}(y=1|\mathbf{x})}{\mathbb{P}(y=0|\mathbf{x})}$ is often known as the *odds*. If we know the odds, we know whether $\mathbb{P}(y = 1|\mathbf{x})$ or $\mathbb{P}(y = 0|\mathbf{x})$ is more likely, and we can decide accordingly. Hence, our goal is to determine what are the odds based on $\mathbf{x}$. Arguably, the simplest, most natural approach is to model the odds as a linear combination of the entries in $\mathbf{x}$, i.e.,

$$\frac{\mathbb{P}(y = 1|\mathbf{x})}{\mathbb{P}(y = 0|\mathbf{x})} \;=\; \theta_0 + \theta_1 \mathrm{x}_1 + \theta_2 \mathrm{x}_2 + \cdots + \theta_{\mathrm{D}} \mathrm{x}_{\mathrm{D}}. \tag{6.3}$$

Redefining $\mathbf{x} = [1 \;\; \mathrm{x}_1 \;\; \mathrm{x}_2 \;\; \cdots \;\; \mathrm{x}_{\mathrm{D}}]^\mathsf{T}$ and letting $\boldsymbol{\theta} = [\theta_0 \;\; \theta_1 \;\; \theta_2 \;\; \cdots \;\; \theta_{\mathrm{D}}]^\mathsf{T}$ we can rewrite (6.3) as $\boldsymbol{\theta}^\mathsf{T}\mathbf{x}$. The problem with (6.3) is that $\frac{\mathbb{P}(y=1|\mathbf{x})}{\mathbb{P}(y=0|\mathbf{x})} \geq 0$, while $\boldsymbol{\theta}^\mathsf{T}\mathbf{x} \in \mathbb{R}$ (i.e., $\boldsymbol{\theta}^\mathsf{T}\mathbf{x}$ could be negative). To avoid this discrepancy, rather than modeling the odds as a linear combination of $\mathbf{x}$ as in (6.3), logistic regression instead uses the so-called *log-odds*, obtained by applying the log function to the odds:

$$\log\left(\frac{\mathbb{P}(y = 1|\mathbf{x})}{\mathbb{P}(y = 0|\mathbf{x})}\right) \;=\; \boldsymbol{\theta}^\mathsf{T}\mathbf{x}. \tag{6.4}$$

It is from this idea that logistic regression obtains its name. Notice that in (6.4), both the log-odds and $\boldsymbol{\theta}^\mathsf{T}\mathbf{x}$ are real numbers, so there is no longer any discrepancy. Letting $\mathrm{p} := \mathbb{P}(y = 1|\mathbf{x})$ we can rewrite (6.4) as

$$\frac{\mathrm{p}}{1 - \mathrm{p}} \;=\; e^{\boldsymbol{\theta}^\mathsf{T}\mathbf{x}},$$

and solving for p we have:

$$\begin{aligned} \mathrm{p} &= (1 - \mathrm{p})e^{\boldsymbol{\theta}^\mathsf{T}\mathbf{x}}, \\ \mathrm{p} &= e^{\boldsymbol{\theta}^\mathsf{T}\mathbf{x}} - \mathrm{p}(e^{\boldsymbol{\theta}^\mathsf{T}\mathbf{x}}), \\ \mathrm{p}(1 + e^{\boldsymbol{\theta}^\mathsf{T}\mathbf{x}}) &= e^{\boldsymbol{\theta}^\mathsf{T}\mathbf{x}}, \\ \mathrm{p} &= \frac{e^{\boldsymbol{\theta}^\mathsf{T}\mathbf{x}}}{1 + e^{\boldsymbol{\theta}^\mathsf{T}\mathbf{x}}}, \end{aligned}$$

which we can further simplify to:

$$\mathrm{p} \;=\; \frac{e^{\boldsymbol{\theta}^\mathsf{T}\mathbf{x}}}{1 + e^{\boldsymbol{\theta}^\mathsf{T}\mathbf{x}}} \;=\; \frac{\frac{e^{\boldsymbol{\theta}^\mathsf{T}\mathbf{x}}}{e^{\boldsymbol{\theta}^\mathsf{T}\mathbf{x}}}}{\frac{1+e^{\boldsymbol{\theta}^\mathsf{T}\mathbf{x}}}{e^{\boldsymbol{\theta}^\mathsf{T}\mathbf{x}}}} \;=\; \frac{1}{\frac{1}{e^{\boldsymbol{\theta}^\mathsf{T}\mathbf{x}}} + \frac{e^{\boldsymbol{\theta}^\mathsf{T}\mathbf{x}}}{e^{\boldsymbol{\theta}^\mathsf{T}\mathbf{x}}}} \;=\; \frac{1}{\frac{1}{e^{\boldsymbol{\theta}^\mathsf{T}\mathbf{x}}} + 1} \;=\; \frac{1}{1 + e^{-\boldsymbol{\theta}^\mathsf{T}\mathbf{x}}}. \tag{6.5}$$

To summarize, logistic regression is modeling

$$\mathbb{P}(y = 1|\mathbf{x}) \;=\; \frac{1}{1 + e^{-\boldsymbol{\theta}^\mathsf{T}\mathbf{x}}}. \tag{6.6}$$

The right hand side of (6.6) is often called *logistic* function. It follows that we can rewrite our decision function (6.2) as:

$$f(\mathbf{x}) \;=\; \begin{cases} 1 & \text{if } \frac{1}{1+e^{-\boldsymbol{\theta}^\mathsf{T}\mathbf{x}}} > 1 - \frac{1}{1+e^{-\boldsymbol{\theta}^\mathsf{T}\mathbf{x}}}, \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, since $\mathbb{P}(y = 1|\mathbf{x}) + \mathbb{P}(y = 0|\mathbf{x}) = 1$, we can further simplify this as

$$f(\mathbf{x}) \;=\; \begin{cases} 1 & \text{if } \frac{1}{1+e^{-\boldsymbol{\theta}^{\mathsf{T}}\mathbf{x}}} > \frac{1}{2}, \\ 0 & \text{otherwise.} \end{cases} \tag{6.7}$$

This means that if you want to know whether you would have survived, all you have to do is plug your feature vector $\mathbf{x}$ in (6.7), and decide accordingly. The catch here is that (6.7) depends on $\boldsymbol{\theta}$, which you do not know a priori. So, which $\boldsymbol{\theta}$ should you use? The answer is: you have to learn it.

6.3 Learning θ

Logistic regression uses (6.7) to decide whether $y = 0$ or $y = 1$ based on $\mathbf{x}$. However, our function f in (6.7) depends on $\boldsymbol{\theta}$, which is unknown a priori. To learn $\boldsymbol{\theta}$ we use *training* data, meaning a collection of feature vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{\mathrm{N}}$ and their corresponding *labels* $\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_{\mathrm{N}}$. In our example, this would mean feature vectors as in (6.1) from N people, and their labels indicating whether they survived or not.

In words, our goal is to find the parameter $\boldsymbol{\theta}$ that best explains our training samples. By doing so we are effectively finding the *best target function* f in the *family* of functions that have the form in (6.7). In this case, by *best* we mean the function f that maximizes the *likelihood* of our sample. In supervised learning jargon, the likelihood is the *cost function* that we are trying to optimize. Intuitively, this likelihood is the chance that our observed samples are correctly predicted by f, i.e., the probability that $\mathrm{y}_{\mathrm{i}} = f(\mathbf{x}_{\mathrm{i}})$, for every $\mathrm{i} = 1, \ldots, \mathrm{N}$.

6.4 Likelihood

Recall that a probability distribution $\mathbb{P}(y|\boldsymbol{\theta})$ determines the frequency with which that a random variable y takes each value, given some parameter $\boldsymbol{\theta}$. For example, if $y \sim \text{Bernoulli}(\theta)$, with $\theta = 1/2$, then the probability that y takes the value 1 is $\mathbb{P}(y = 1|\theta) = \theta = 1/2$.

Conversely, the *likelihood* $\mathbb{P}(\mathrm{y}|\boldsymbol{\theta})$ determines the probability that a parameter $\boldsymbol{\theta}$ was the one that generated a sample y. We emphasize this distinction using y instead of y, to indicate that y is already known, i.e., observed data that has already taken a specific value. Under the same Bernoulli example, if we observe $\mathrm{y} = 1$, then the likelihood of the parameter θ is $\mathbb{P}(\mathrm{y} = 1|\theta) \;=\; \theta$.

The probability and the likelihood may *look* a lot alike. The difference is very subtle, and mainly conceptually: the probability $\mathbb{P}(y|\boldsymbol{\theta})$ is a function where y is the variable, and $\boldsymbol{\theta}$ is fixed. In contrast, the likelihood $\mathbb{P}(\mathrm{y}|\boldsymbol{\theta})$ is a function where $\boldsymbol{\theta}$ is the variable, and y is fixed. We use $\mathbb{P}(y|\boldsymbol{\theta})$ when we know $\boldsymbol{\theta}$ and want to guess y; we use $\mathbb{P}(\mathrm{y}|\boldsymbol{\theta})$ when we have already observed data with the specific value y, and we want to guess the parameter $\boldsymbol{\theta}$ that generated it.

Example 6.1. Suppose $y_1, \ldots, y_6$ are *independently and identically distributed* (i.i.d.) according to a

Bernoulli($1/4$) distribution. Then the probability that $y_1 = y_2 = y_3 = 1$, and $y_4 = y_5 = y_6 = 0$ is:

$$\begin{aligned}\mathbb{P}(y_1 = y_2 = y_3 = 1, y_4 = y_5 = y_6 = 0|\theta) &= \prod_{i=1}^{3}\mathbb{P}(y_i = 1|\theta) \cdot \prod_{i=4}^{6}\mathbb{P}(y_i = 0|\theta) \\ &= \theta^3(1-\theta)^3 = (1/4)^3\ (3/4)^3.\end{aligned}$$

Instead, suppose that we observe $y_1 = y_2 = y_3 = 1$, and $y_4 = y_5 = y_6 = 0$. Then the likelihood of θ under this sample is:

$$\begin{aligned}\mathbb{P}(y_1 = y_2 = y_3 = 1, y_4 = y_5 = y_6 = 0|\theta) &= \prod_{i=1}^{3}\mathbb{P}(y_i = 1|\theta) \cdot \prod_{i=4}^{6}\mathbb{P}(y_i = 0|\theta) \\ &= \theta^3(1-\theta)^3.\end{aligned}$$

Based on this sample, which would be your intuitive best guess at the value of θ? Is this the same value that maximizes the likelihood $\mathbb{P}(y_1, \ldots, y_6|\theta)$?

6.5 Maximum Likelihood

Back to logistic regression, we can model our training data $y_1, \ldots, y_N$ as i.i.d. realizations of a *Bernoulli*(p) random variable, where $p = \frac{1}{1+e^{-\theta^T x_i}}$. A little thought shows that the likelihood of a *Bernoulli*(p) random variable can be written as:

$$\mathbb{P}(y|p) = p^y(1-p)^{1-y}.$$

Make sure you understand why this is true. By independence, the likelihood of our training sample is:

$$\mathbb{P}(y_1, \ldots, y_N|p) = \prod_{i=1}^{N}\mathbb{P}(y_i|p) = \prod_{i=1}^{N} p^{y_i}(1-p)^{1-y_i}.$$

Since p is in turn a function of the unknown parameter $\boldsymbol{\theta}$ and the known data $\mathbf{x}_1, \ldots, \mathbf{x}_N$, we can rewrite this likelihood as

$$\begin{aligned}\mathbb{P}(y_1, \ldots, y_N|\mathbf{x}_1, \ldots, \mathbf{x}_N, \boldsymbol{\theta}) &= \prod_{i=1}^{N}\left(\frac{1}{1+e^{-\boldsymbol{\theta}^T\mathbf{x}_i}}\right)^{y_i}\left(1 - \frac{1}{1+e^{-\boldsymbol{\theta}^T\mathbf{x}_i}}\right)^{1-y_i} \\ &= \prod_{i=1}^{N}\left(\frac{1}{1+e^{-\boldsymbol{\theta}^T\mathbf{x}_i}}\right)^{y_i}\left(\frac{1}{1+e^{\boldsymbol{\theta}^T\mathbf{x}_i}}\right)^{1-y_i}, \qquad (6.8)\end{aligned}$$

where the last step follows by similar manipulations as in (6.5). To ease our notation we will use $\mathbb{P}(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$ as shorthand for $\mathbb{P}(y_1, \ldots, y_N|\mathbf{x}_1, \ldots, \mathbf{x}_N, \boldsymbol{\theta})$. Our goal is to find the $\boldsymbol{\theta}$ that maximizes this likelihood. Maximizing products as in (6.8) can be difficult (as you know from the chain rule of derivatives), so to simplify this maximization, we will use a common trick: apply log, so that products transform into sums, which are easily maximized (because of the linearity of derivatives: the derivative of a sum is the sum of derivatives). We know we can do this because $\mathbb{P}$ is positive (so we can apply log), and log is monotonically increasing, implying that

$$\arg\max_{\boldsymbol{\theta}\in\mathbb{R}^{D+1}}\ \mathbb{P}(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}\in\mathbb{R}^{D+1}}\ \log\left[\mathbb{P}(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})\right].$$

So instead of maximizing the likelihood directly, we can equivalently maximize the so-called log-likelihood:

$$\begin{aligned}
\ell(\boldsymbol{\theta}) \;&:=\; \log\left[\mathbb{P}(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})\right] \\
&=\; \log\left[\prod_{\mathrm{i}=1}^{\mathrm{N}}\left(\frac{1}{1+e^{-\boldsymbol{\theta}^{\mathsf{T}}\mathbf{x}_{\mathrm{i}}}}\right)^{\mathrm{y}_{\mathrm{i}}}\left(\frac{1}{1+e^{\boldsymbol{\theta}^{\mathsf{T}}\mathbf{x}_{\mathrm{i}}}}\right)^{1-\mathrm{y}_{\mathrm{i}}}\right] \\
&=\; \sum_{\mathrm{i}=1}^{\mathrm{N}}\log\left[\left(\frac{1}{1+e^{-\boldsymbol{\theta}^{\mathsf{T}}\mathbf{x}_{\mathrm{i}}}}\right)^{\mathrm{y}_{\mathrm{i}}}\left(\frac{1}{1+e^{\boldsymbol{\theta}^{\mathsf{T}}\mathbf{x}_{\mathrm{i}}}}\right)^{1-\mathrm{y}_{\mathrm{i}}}\right] \\
&=\; \sum_{\mathrm{i}=1}^{\mathrm{N}}\left[\mathrm{y}_{\mathrm{i}}\log\left(\frac{1}{1+e^{-\boldsymbol{\theta}^{\mathsf{T}}\mathbf{x}_{\mathrm{i}}}}\right)+(1-\mathrm{y}_{\mathrm{i}})\log\left(\frac{1}{1+e^{\boldsymbol{\theta}^{\mathsf{T}}\mathbf{x}_{\mathrm{i}}}}\right)\right], \qquad (6.9)
\end{aligned}$$

which is easier to maximize than (6.8) because it contains a sum, rather than a product. Sadly, (6.9) is still complex enough that it cannot be maximized with our calculus 101 recipe (take derivative, set to zero, and solve for the optimizer). Instead we can use something like gradient ascent, which can be summarized as follows:

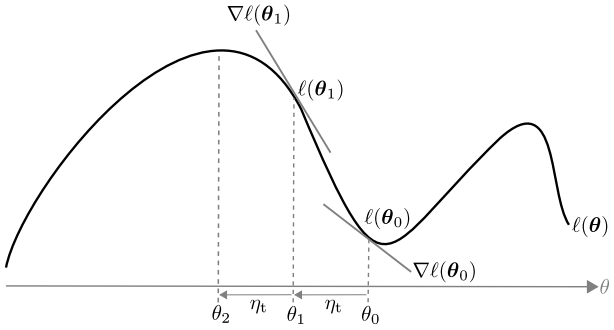
- Find an expression for the gradient of the function you want to optimize. In our case it would be:

$$\nabla\ell(\boldsymbol{\theta}) \;:=\; \frac{\mathrm{d}\ell(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} \;=\; \sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\mathrm{y}_{\mathrm{i}}-\frac{1}{1+e^{-\boldsymbol{\theta}^{\mathsf{T}}\mathbf{x}_{\mathrm{i}}}}\right)\mathbf{x}_{\mathrm{i}}.$$

- Initialize $\boldsymbol{\theta}_0$ with your best guess.
- Repeat until convergence:

$$\boldsymbol{\theta}_{\mathrm{t}+1} \;=\; \boldsymbol{\theta}_{\mathrm{t}} \;+\; \eta_{\mathrm{t}}\nabla\ell(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathrm{t}}},$$

  where η_{t} is the step-size parameter at time t:



In general, gradient descent may get stuck in a local optima of the objective function its trying to optimize. Fortunately, the objective function in logistic regression is a log-likelihood from the exponential family, which as the next theorem shows, is concave.

Theorem 6.1 (Concavity of the log-likelihood)**.** The log-likelihood function

$$\ell(\boldsymbol{\theta}) \;=\; \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{y_i} \log\left(\frac{1}{1+e^{-\boldsymbol{\theta}^\mathsf{T}\mathbf{x}_\mathrm{i}}}\right) + (1-\mathrm{y_i}) \log\left(\frac{1}{1+e^{\boldsymbol{\theta}^\mathsf{T}\mathbf{x}_\mathrm{i}}}\right)$$

is concave in $\boldsymbol{\theta}$.

Proof. Follows directly because the Hessian is negative semidefinite:

$$\frac{\mathrm{d}^2\ell(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}^2} \;=\; -\sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{e^{-\boldsymbol{\theta}^\mathsf{T}\mathbf{x}_\mathrm{i}}}{(1+e^{-\boldsymbol{\theta}^\mathsf{T}\mathbf{x}_\mathrm{i}})^2} \mathbf{x}_\mathrm{i}\mathbf{x}_\mathrm{i}^\mathsf{T}.$$

We know this because matrices of the form $\mathbf{x}_\mathrm{i}\mathbf{x}_\mathrm{i}^\mathsf{T}$ are positive semidefinite and $\frac{e^{-\boldsymbol{\theta}^\mathsf{T}\mathbf{x}_\mathrm{i}}}{(1+e^{-\boldsymbol{\theta}^\mathsf{T}\mathbf{x}_\mathrm{i}})^2} \geq 0$, and the sum of positive semidefinite matrices scaled by non-negative numbers is also positive semidefinite.

□

Theorem 6.1 ensures that gradient descent will find (or get arbitrarily close to) the *maximum likelihood estimator* (MLE):

$$\hat{\boldsymbol{\theta}} \;:=\; \underset{\boldsymbol{\theta}\in\mathbb{R}^{\mathrm{D}+1}}{\arg\max}\; \mathbb{P}(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}).$$

Once we have found $\hat{\boldsymbol{\theta}}$, we can determine whether $y = 0$ or $y = 1$ for a new sample with features $\mathbf{x}$, by simply using (6.7), with $\hat{\boldsymbol{\theta}}$ instead of $\boldsymbol{\theta}$.

6.6 Asymptotic Confidence

At this point we are capable of learning the best parameter $\boldsymbol{\theta}$ (which in turn determines the best function of the form in (6.7)) that explains our training data, and use it to predict new y's. The next question we should be asking is: how much can I trust my prediction? Perhaps we do not care too much about this for our hypothetical Titanic survival example. But what about other cases where decisions are crucial, for example in medical applications. Or going back to our edible vs. poisonous mushroom example, would you eat a mushroom that your logistic regression algorithm classifies as edible?

In linear regression we addressed this question using confidence intervals: we showed that our estimate

$$\hat{\boldsymbol{\theta}} \;=\; (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} \tag{6.10}$$

was distributed $\mathcal{N}(\boldsymbol{\theta}^\star, \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1})$, and computed the likelihood that it was too far from the true $\boldsymbol{\theta}^\star$, which in turn determined the reliability of our prediction $\hat{\mathrm{y}} = \mathbf{x}^\mathsf{T}\hat{\boldsymbol{\theta}}$.

Unfortunately, since maximizing the likelihood in logistic regression involves a numeric algorithm (like gradient ascent), we don't have a closed-form expression for $\hat{\boldsymbol{\theta}}$ like (6.10) in linear regression. Fortunately, since $\hat{\boldsymbol{\theta}}$ is an MLE, we can use the next theorem, stating that estimators of this type follow an asymptotic Normal distribution as the sample size N grows.

Definition 6.1 (Convergence in Distribution)**.** Let Z_{N}, Z be random variables with cumulative density functions F_{N} and F. We say Z_{N} *converges in distribution* to Z, denoted as $Z_{\mathrm{N}} \xrightarrow{d} Z$, if

$$\lim_{\mathrm{N}\to\infty} F_{\mathrm{N}}(z) = F(z) \qquad \text{for all } z \text{ in which } F \text{ is continuous.}$$

Intuitively, $Z_{\mathrm{N}} \xrightarrow{d} Z$ means that as N grows, the distribution of Z_{N} is approximately the same as the distribution of Z. Notice that our estimator $\hat{\boldsymbol{\theta}}$ is a random vector that depends on the number of samples N. The next theorem shows that as N grows, $\hat{\boldsymbol{\theta}}$ (playing the role of Z_{N}) converges in distribution to a Normal random variable.

Theorem 6.2 (Asymptotic distribution of the MLE)**.** Let $y_1, \ldots, y_{\mathrm{N}}$ be independent random variables, where $y_{\mathrm{i}} \sim \mathbb{P}(y|\boldsymbol{x}_{\mathrm{i}}, \boldsymbol{\theta}^\star)$, with $\boldsymbol{\theta}^\star \in \mathbb{R}^{\mathrm{D}+1}$. Define

$$\ell(\boldsymbol{\theta}) := \sum_{\mathrm{i}=1}^{\mathrm{N}} \log \mathbb{P}(\mathrm{y_i}|\mathbf{x}_{\mathrm{i}}, \boldsymbol{\theta}), \qquad \text{and} \qquad \hat{\boldsymbol{\theta}} := \operatorname*{arg\,max}_{\boldsymbol{\theta}\in\mathbb{R}^{\mathrm{D}+1}} \ell(\boldsymbol{\theta}).$$

Suppose $\dfrac{\mathrm{d}^2\ell(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}^2}$ exists. Then as $\mathrm{N}\to\infty$,

$$\hat{\boldsymbol{\theta}} \xrightarrow{d} \mathcal{N}\Big(\boldsymbol{\theta}^\star, \mathbf{I}_{\boldsymbol{\theta}^\star}^{-1}\Big),$$

where $\mathbf{I}_{\boldsymbol{\theta}^\star}$ is the *Fisher-information matrix*, defined as:

$$\mathbf{I}_{\boldsymbol{\theta}^\star} := -\mathbb{E}\left[\frac{\mathrm{d}^2\ell(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}^2}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\star}\right].$$

Proof. The proof follows as a consequence of the central limit theorem, which says that if $z_1, \ldots, z_{\mathrm{N}}$ are i.i.d. with mean μ and variance σ^2, then $\frac{\sqrt{\mathrm{N}}}{\sigma}\left(\sum_{\mathrm{i}=1}^{\mathrm{N}} z_{\mathrm{i}} - \mu\right)$ is asymptotically distributed $\mathcal{N}(0,1)$. For a detailed proof see Theorem 10.1.12 in *Statistical Inference* by George Casella and Roger L. Berger, second edition, or Theorem 4.17 in *Mathematical Statistics* by Jun Shao, second edition. □

Using Theorem 6.2 we can approximate the distribution of $\hat{\boldsymbol{\theta}}$. All we need to do is compute the Fisher-information matrix, given by

$$\mathbf{I}_{\boldsymbol{\theta}^\star} := -\mathbb{E}\left[\frac{\mathrm{d}^2\ell(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}^2}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\star}\right] = \mathbb{E}\left[\sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{e^{-\boldsymbol{\theta}^\mathsf{T}\mathbf{x}_{\mathrm{i}}}}{(1+e^{-\boldsymbol{\theta}^\mathsf{T}\mathbf{x}_{\mathrm{i}}})^2}\mathbf{x}_{\mathrm{i}}\mathbf{x}_{\mathrm{i}}^\mathsf{T}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\star}\right] = \sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{e^{-\boldsymbol{\theta}^{\star\mathsf{T}}\mathbf{x}_{\mathrm{i}}}}{(1+e^{-\boldsymbol{\theta}^{\star\mathsf{T}}\mathbf{x}_{\mathrm{i}}})^2}\mathbf{x}_{\mathrm{i}}\mathbf{x}_{\mathrm{i}}^\mathsf{T}.$$

At this point we can derive confidence intervals for the log-odds, similar to what we did in linear regression. To see this, recall that the the log-odds are defined as

$$\omega^\star := \log\left(\frac{\mathbb{P}(y=1|\mathbf{x})}{\mathbb{P}(y=0|\mathbf{x})}\right) = \boldsymbol{\theta}^{\star\mathsf{T}}\mathbf{x}.$$

By the invariance property of the MLE, we know that the MLE of $\omega^\star$ is given by

$$\hat{\omega} \;:=\; \hat{\boldsymbol{\theta}}^\mathsf{T}\mathbf{x},$$

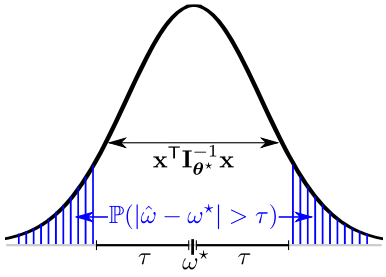
where $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}^\star$. By Theorem 6.2, we know that $\hat{\boldsymbol{\theta}} \xrightarrow{d} \mathcal{N}\Big(\boldsymbol{\theta}^\star, \mathbf{I}_{\boldsymbol{\theta}^\star}^{-1}\Big)$, which further implies $\hat{\omega} \xrightarrow{d} \mathcal{N}\left(\boldsymbol{\theta}^{\star\mathsf{T}}\mathbf{x}, \mathbf{x}^\mathsf{T}\mathbf{I}_{\boldsymbol{\theta}^\star}^{-1}\mathbf{x}\right)$, or equivalently

$$\hat{\omega} \xrightarrow{d} \mathcal{N}\left(\omega^\star, \mathbf{x}^\mathsf{T}\mathbf{I}_{\boldsymbol{\theta}^\star}^{-1}\mathbf{x}\right).$$

It follows that the asymptotic probability that $\hat{\omega}$ is τ-away from the true $\omega^\star$ is

$$\mathbb{P}(|\hat{\omega} - \omega^\star| > \tau) \;=\; 2\;\Phi_{\mathcal{N}}\left(\;\tau\;\middle|\;0, \mathbf{x}^\mathsf{T}\mathbf{I}_{\boldsymbol{\theta}^\star}^{-1}\mathbf{x}\right),$$

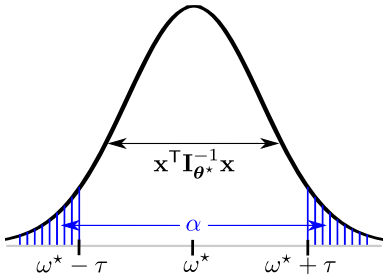
where $\Phi_{\mathcal{N}}(\;\cdot\;|\;\mu, \nu)$ is the tail function of the $\mathcal{N}(\mu, \nu)$ distribution:



Conversely, given a desired significance level α (typically set to 0.05), we can find a threshold

$$\tau \;=\; \Phi_{\mathcal{N}}^{-1}\left(\;\alpha/2\;\middle|\;0, \mathbf{x}^\mathsf{T}\mathbf{I}_{\boldsymbol{\theta}^\star}^{-1}\mathbf{x}\right)$$

such that $\mathbb{P}(|\hat{\omega} - \omega^\star| \leq \tau) = 1 - \alpha$:



Equivalently, we conclude that with probability $1 - \alpha$ (typically set to 0.95), the *true* (but unknown) $\omega^\star$ is in the *confidence interval* $(\hat{\omega} - \tau, \hat{\omega} + \tau)$.

6.7 Learning Significant Features

Another cool thing about Theorem 6.2 is that now that we know the asymptotic distribution of $\hat{\boldsymbol{\theta}}$, we can also identify its significant features using a similar hypothesis test as in linear regression:

$$\begin{aligned} H_0:\; \hat{\theta}_\mathrm{j} \;&\sim\; \mathcal{N}(\theta_\mathrm{j}^\star, \nu_\mathrm{j}^2), \quad \theta_\mathrm{j}^\star = 0 \qquad &&\Rightarrow \mathrm{j}^{\mathrm{th}} \text{ feature is irrelevant,} \\ H_1:\; \hat{\theta}_\mathrm{j} \;&\sim\; \mathcal{N}(\theta_\mathrm{j}^\star, \nu_\mathrm{j}^2), \quad \theta_\mathrm{j}^\star \neq 0 \qquad &&\Rightarrow \mathrm{j}^{\mathrm{th}} \text{ feature is significant.} \end{aligned}$$

where ν_{j}^2 denotes the $(\mathrm{j},\mathrm{j})^{\text{th}}$ entry in the covariance matrix $\mathbf{I}_{\boldsymbol{\theta}^\star}^{-1}$. Recall that the main idea to address such hypothesis test problem is to use a likelihood ratio test (LRT):

$$\frac{\mathbb{P}(\hat{\theta}_{\mathrm{j}}|\theta_{\mathrm{j}}^\star \neq 0)}{\mathbb{P}(\hat{\theta}_{\mathrm{j}}|\theta_{\mathrm{j}}^\star = 0)} \underset{H_0}{\overset{H_1}{\gtrless}} 1. \tag{6.11}$$

Here $\mathbb{P}(\hat{\theta}_{\mathrm{j}}|\theta_{\mathrm{j}}^\star \neq 0)$ denotes the likelihood of $\hat{\theta}_{\mathrm{j}}$ under H_1, and similarly $\mathbb{P}(\hat{\theta}_{\mathrm{j}}|\theta_{\mathrm{j}}^\star = 0)$ denotes the likelihood of $\hat{\theta}_{\mathrm{j}}$ under H_0. In words, (6.11) decides H_1 if the *likelihood ratio* $\Lambda(\hat{\theta}_{\mathrm{j}}) := \frac{\mathbb{P}(\hat{\theta}_{\mathrm{j}}|\theta_{\mathrm{j}}^\star \neq 0)}{\mathbb{P}(\hat{\theta}_{\mathrm{j}}|\theta_{\mathrm{j}}^\star \neq 0)}$ is larger than 1 (meaning the likelihood under H_1 is larger than under H_0), and decides H_0 otherwise. However, since we do not know the specific value of $\theta_{\mathrm{j}}^\star$ under H_1, we cannot compute (6.11) directly. Instead we have to use a *generalized likelihood ratio test* (GLRT):

$$\frac{\max\limits_{\theta_{\mathrm{j}}^\star \neq 0} \mathbb{P}(\hat{\theta}_{\mathrm{j}}|\theta_{\mathrm{j}}^\star)}{\mathbb{P}(\hat{\theta}_{\mathrm{j}}|\theta_{\mathrm{j}}^\star = 0)} \underset{H_0}{\overset{H_1}{\gtrless}} \tau, \tag{6.12}$$

where τ can be chosen to bound the probability of a certain type of error (e.g., deciding H_1 when H_0 is true, often called Type 1 error) or guarantee the probability of a correct decision (e.g., correctly rejecting H_0). The main idea behind (6.12) is to substitute $\theta_{\mathrm{j}}^\star$ with its MLE, which we already know from before to be $\hat{\theta}_{\mathrm{j}}$. Then our test becomes:

$$\Lambda(\hat{\theta}_{\mathrm{j}}) \;=\; \frac{\mathbb{P}(\hat{\theta}_{\mathrm{j}}|\theta_{\mathrm{j}}^\star = \hat{\theta}_{\mathrm{j}})}{\mathbb{P}(\hat{\theta}_{\mathrm{j}}|\theta_{\mathrm{j}}^\star = 0)} \;=\; \frac{\frac{1}{\sqrt{2\pi}\nu_{\mathrm{j}}} e^{-\frac{1}{2}\left(\frac{\hat{\theta}_{\mathrm{j}}-\hat{\theta}_{\mathrm{j}}}{\nu_{\mathrm{j}}}\right)^2}}{\frac{1}{\sqrt{2\pi}\nu_{\mathrm{j}}} e^{-\frac{1}{2}\left(\frac{\hat{\theta}_{\mathrm{j}}-0}{\nu_{\mathrm{j}}}\right)^2}} \;=\; e^{\frac{\hat{\theta}_{\mathrm{j}}^2}{2\nu_{\mathrm{j}}^2}} \underset{H_0}{\overset{H_1}{\gtrless}} \tau.$$

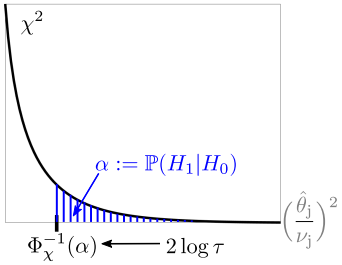
Taking log and with minor algebra manipulations we can further simplify our test into:

$$\left(\frac{\hat{\theta}_{\mathrm{j}}}{\nu_{\mathrm{j}}}\right)^2 \underset{H_0}{\overset{H_1}{\gtrless}} 2\log\tau.$$

Under H_0, $\hat{\theta}_{\mathrm{j}} \sim \mathcal{N}(0, \nu_{\mathrm{j}}^2)$, so $\hat{\theta}_{\mathrm{j}}/\nu_{\mathrm{j}} \sim \mathcal{N}(0,1)$, which implies $(\hat{\theta}_{\mathrm{j}}/\nu_{\mathrm{j}})^2 \sim \chi^2$. Given a desired significance level α (probability of deciding H_1 given that H_0 is true, typically set to 0.05), we can select τ as:

$$\tau \;=\; e^{\frac{1}{2}\Phi_\chi^{-1}(\alpha)},$$

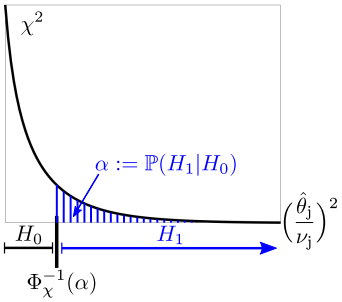
where Φ_χ is the tail function of the χ^2 distribution:



With this, our test further simplifies into:

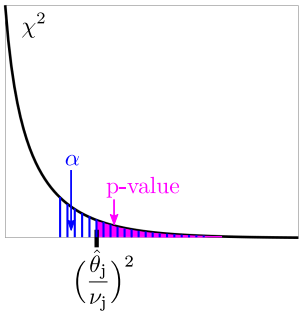
$$\left(\frac{\hat{\theta}_j}{\nu_j}\right)^2 \underset{H_0}{\overset{H_1}{\gtrless}} \Phi_\chi^{-1}(\alpha) \tag{6.13}$$

In words, (6.13) says: decide H_1 if $\hat{\theta}_j^2 > \nu_j^2 \Phi_\chi^{-1}(\alpha)$, and decide H_0 otherwise, which matches our intuition, essentially saying: if $|\hat{\theta}_j|$ is large enough, conclude that $\theta_j^\star \neq 0$, and that the j^{th} feature is significant; conversely, if $|\hat{\theta}_j|$ is too small, conclude that $\theta_j^\star = 0$, and that the j^{th} feature is irrelevant:



Finally, given an instance of the test statistic $(\hat{\theta}_j/\nu_j)^2$, its p-value (indicating the probability of observing a larger test statistic under H_0) is

$$\Phi_\chi\left(\frac{\hat{\theta}_j^2}{\nu_j^2}\right).$$



6.8 Logistic Regression Recipe

To summarize, here is how we would use logistic regression in a typical scenario:

- Collect *labels* $\mathbf{y} \in \{0,1\}^{\mathrm{N}}$ and *features* $\mathbf{X} \in \mathbb{R}^{\mathrm{N}\times\mathrm{D}}$ from N samples.
- Find the MLE $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\mathbb{R}^{\mathrm{D}+1}} \mathbb{P}(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})$ using gradient descent or your favorite method.
- Given a new sample $\mathbf{x}$, decide:

$$\hat{y} \;=\; \begin{cases} 1 & \text{if } \frac{1}{1+e^{-\hat{\boldsymbol{\theta}}^{\mathsf{T}}\mathbf{x}}} > \frac{1}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

- Given a significance level α (typically set to 0.05), we know that with probability $1-\alpha$, the *true* log-odds $\omega^\star$ lie in the confidence interval $(\hat{\omega}-\tau, \hat{\omega}+\tau)$, where $\hat{\omega} = \hat{\boldsymbol{\theta}}^{\mathsf{T}}\mathbf{x}$, and $\tau = \Phi_{\mathcal{N}}^{-1}\big(\,\alpha/2 \;\big|\; 0, \mathbf{x}^{\mathsf{T}}\mathbf{I}_{\hat{\boldsymbol{\theta}}}^{-1}\mathbf{x}\big)$, and $\Phi_{\mathcal{N}}(\,\cdot \mid \mu,\nu)$ is the tail function of the $\mathcal{N}(\mu,\nu)$ distribution. Notice that since $\boldsymbol{\theta}^\star$ is unknown (and hence so is $\mathbf{I}_{\boldsymbol{\theta}^\star}^{-1}$), here we are substituting $\mathbf{I}_{\boldsymbol{\theta}^\star}^{-1}$ with $\mathbf{I}_{\hat{\boldsymbol{\theta}}}^{-1}$, which by the invariance property is also an MLE.
- With p-value $\Phi_\chi(\hat{\theta}_{\mathrm{j}}^2/\nu_{\mathrm{j}}^2)$, conclude that the j^{th} feature is significant if $\hat{\theta}_{\mathrm{j}}^2 > \nu_{\mathrm{j}}^2\Phi_\chi^{-1}(\alpha)$, and decide it is irrelevant otherwise; here ν_{j}^2 is the $(\mathrm{j},\mathrm{j})^{\mathrm{th}}$ entry in $\mathsf{cov}(\hat{\boldsymbol{\theta}}) = \mathbf{I}_{\hat{\boldsymbol{\theta}}}^{-1}$. Here Φ_χ is the tail function of the χ^2 distribution.