Indian Institute of Technology, Kanpur
Probabilistic Machine Learning and Inference

Instructors: Piyush Rai
Authors: Gurpreet Singh
Email: guggu@iitk.ac.in

SCRIBE

4

Bayesian Approach to Logistic Regression

1. Probabilistic Models for Classification

Our goal is to learn $\mathbb{P}\left[\mathbf{y} \,\middle|\, X\right]$, which in case of classification will be a discrete distribution, precisely a Bernoulli or a Multinoulli. There are usually two approaches to learn $\mathbb{P}\left[\mathbf{y} \,\middle|\, X\right]$, *i.e.* Discriminative and Generative

1. **Discriminative Classification** Discriminative models learn the (hard or soft) boundary between classes. In this type of classification, we directly model $\mathbb{P}\left[\mathbf{y} \,\middle|\, X\right]$ and do not model the distribution of the inputs X.
2. **Generative Classification** Generative models learn the distribution of individual classes, as along with the marginal likelihood, we also learn the distribution of the inputs. In this case, we model $\mathbb{P}\left[y \,\middle|\, x\right]$ indirectly, using the bayes rule, *i.e.* $\mathbb{P}\left[\mathbf{y} \,\middle|\, X\right] = \frac{\mathbb{P}[\mathbf{y}]\mathbb{P}\left[X \,\middle|\, \mathbf{y}\right]}{\mathbb{P}[X]}$

   Hence, this approach first requires learning the class marginal ($\mathbb{P}[\mathbf{y}]$) and the class-conditional ($\mathbb{P}\left[X \,\middle|\, \mathbf{y}\right]$). This is usually harder than discriminative classification but it also has other advantanges, such as it is possible to learn the input data itself, and can be used even in case of incomplete data.

Both, generative and discriminative approaches can be learned using Bayesian as well as non-Bayesian Inference techniques. In this scribe, we only discuss Logisitic Regression.

2. Logistic Regression: Point Estimate

Logisitic Regression is an example of discriminative binary classification, *i.e.* the output space $\mathcal{Y} = \{0, 1\}$. Logistic Regression models the $\mathcal{X}$ to $\mathcal{Y}$ relationship using the *sigmoid* (σ) function, defined as below

$$\mathbb{P}\left[y = 1 \,\middle|\, \mathbf{x}, \mathbf{w}\right] \quad = \quad \sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}\right) \quad = \quad \frac{1}{1 + \exp\left(-\mathbf{w}^{\mathrm{T}}\mathbf{x}\right)} \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^D$ is the weight vector. We can also write $\mathbb{P}\left[y = 0 \,\middle|\, \mathbf{x}, \mathbf{w}\right] = 1 - \sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}\right)$.

We can also use other functions to map the input space $\mathcal{X} \subseteq \mathbb{R}^D$ to a valid output space for a probability measure, *i.e.* $[0, 1]$. For example, the probit function (Probit Regression) maps the input space using the CDF of the normal distribution

$$\mathbb{P}\left[y = 1 \,\middle|\, \mathbf{x}\right] \quad = \quad \boldsymbol{\phi}(\mathbf{w}^{\mathrm{T}}\mathbf{x})$$

where $\boldsymbol{\phi} = \text{ CDF}\left(\mathcal{N}(0, 1)\right)$

We model the likelihood distribution as a Bernoulli.

$$\mathbb{P}\left[y \,\middle|\, \mathbf{x}, \mathbf{w}\right] \quad = \quad \mu^{\mathbb{I}[y=1]}(1 - \mu)^{\mathbb{I}[y=0]} < ++ > \tag{2}$$

where $\mu = \sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}\right)$

Given N observations (i.i.d.), we can do point estimation for $\mathbf{w}$ by maximizing the likelihood (or minimize negative log-likelihood).

$$\mathbf{w}_{\mathrm{MLE}} \quad = \quad \underset{\mathbf{w}}{\arg\min} \; -\sum_{n=1}^{N} \log\left(\mathbb{P}\left[y_n \,\middle|\, \mathbf{x}_n, \mathbf{w}\right]\right) \tag{3}$$

We cannot minize usign derivatives, as the solution is not in closed form. Therefore we need to use function optimization techniques such as gradient descent or second order methods such as Newton's method. Since this is a convex loss function, we will attain a global minima using function optimization.

It is also possible to regularize $\mathbf{w}$ (MAP estimation) to prevent overfitting.

$$\mathbf{w}_{MAP} \quad = \quad \arg\max_{\mathbf{w}} \sum_{n=1}^{N} \log\left(\mathbb{P}\left[\, y_n \,\middle|\, \mathbf{x}_n, \mathbf{w} \,\right]\right) + \log\left(\mathbb{P}\left[\, \mathbf{w} \,\right]\right) \tag{4}$$

3. Bayesian Logistic Regression

Since MLE / MAP give only a point estimate, they do not give a full inference on the data and the outputs. Therefore, we would like to infer the full posterioir over $\mathbf{w}$

Similar to bayesian linear regression case, we assume a Gaussian prior on $\mathbf{w}$

$$\mathbb{P}\left[\, \mathbf{w} \,\right] \quad = \quad \mathcal{N}\left(0, \lambda^{-1}\mathbf{I}_D\right) \tag{5}$$

Using bayes rule, we have

$$\mathbb{P}\left[\, \mathbf{w} \,\middle|\, X, \mathbf{y} \,\right] \quad \propto \quad \mathbb{P}\left[\, \mathbf{y} \,\middle|\, X, \mathbf{w} \,\right] \, \mathbb{P}\left[\, \mathbf{w} \,\right]$$

From Equations 2 and 5, we have

$$\mathbb{P}\left[\, \mathbf{w} \,\middle|\, X, \mathbf{y} \,\right] \quad \propto \quad \prod_{n=1}^{N} \mu^{\mathbb{I}[\, y_n = 1\,]} (1-\mu)^{\mathbb{I}[\, y_n = 0\,]} \exp\left(-\frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right)$$

where $\mu = \sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}\right)$.

However, the integral of such an expression is intractable. Hence, we can't get a closed form for $\mathbb{P}\left[\, \mathbf{w} \,\middle|\, X, \mathbf{y} \,\right]$. One solution to this is to approximate the posterior. There are several ways to do it, such as MCMC and variational inference (more on this later), and *Laplace Approximation*.

We will discuss Laplace Approximation as a solution to our problem.

3.1 Laplace Approximation

Laplace Approximation is the approximation of an intractable distribution using a gaussian with the mode of the distribution as the mean, and the hessian matrix of the negative log probability of the distribution as the precision matrix.

More formally, we approximate the posterior $\mathbb{P}\left[\, \boldsymbol{\theta} \,\middle|\, \mathcal{D} \,\right] = \frac{\mathbb{P}\left[\, \mathcal{D} \,\middle|\, \boldsymbol{\theta} \,\right]\mathbb{P}\left[\, \boldsymbol{\theta} \,\right]}{\mathbb{P}\left[\, \mathcal{D} \,\right]}$ by th following gaussian

$$\mathbb{P}\left[\, \boldsymbol{\theta} \,\middle|\, \mathcal{D} \,\right] \quad \approx \quad \mathcal{N}\left(\boldsymbol{\theta}_{\mathrm{MAP}}, H^{-1}\right)$$

where

$$\boldsymbol{\theta}_{\mathrm{MAP}} \quad = \quad \arg\max_{\boldsymbol{\theta}} \mathbb{P}\left[\, \boldsymbol{\theta} \,\middle|\, \mathcal{D} \,\right] \quad = \quad \arg\max_{\boldsymbol{\theta}} \mathbb{P}\left[\, \mathcal{D} \,\middle|\, \boldsymbol{\theta} \,\right] \mathbb{P}\left[\, \boldsymbol{\theta} \,\right]$$

and

$$H \quad = \quad -\nabla^2 \left[\, \log\left(\mathbb{P}\left[\, \boldsymbol{\theta} \,\middle|\, \mathcal{D} \,\right]\right) \,\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathrm{MAP}}} \quad = \quad -\nabla^2 \left[\, \log\left(\mathbb{P}\left[\, \boldsymbol{\theta}, \mathcal{D} \,\right]\right) \,\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathrm{MAP}}}$$

The derivation / verification of the Laplace Approximation can be seen using the approximate taylor's expansion. We write the posterior as

$$\begin{aligned}
\mathbb{P}\left[\,\boldsymbol{\theta}\,\middle|\,\mathcal{D}\,\right] &= \frac{\mathbb{P}\left[\,\mathcal{D},\boldsymbol{\theta}\,\right]}{\mathbb{P}\left[\,\mathcal{D}\,\right]} \\
&= \frac{\mathbb{P}\left[\,\mathcal{D},\boldsymbol{\theta}\,\right]}{\int \mathbb{P}\left[\,\mathcal{D},\boldsymbol{\theta}\,\right]\mathrm{d}\boldsymbol{\theta}} \\
&= \frac{\exp\left(\,\log\left(\mathbb{P}\left[\,\mathcal{D},\boldsymbol{\theta}\,\right]\right)\right)}{\int\,\exp\left(\,\log\left(\mathbb{P}\left[\,\mathcal{D},\boldsymbol{\theta}\,\right]\right)\right)\mathrm{d}\boldsymbol{\theta}}
\end{aligned}$$

Suppose $\log\left(\mathbb{P}\left[\,\mathcal{D},\boldsymbol{\theta}\,\right]\right) = f(\boldsymbol{\theta})$, then we can write using 2nd Order Taylor Expansion

$$f(\boldsymbol{\theta}) \quad\approx\quad f(\boldsymbol{\theta}_0) + (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\mathrm{T}} \bigtriangledown f(\theta_0) + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\mathrm{T}} \bigtriangledown^2 f(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$$

Let us choose $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_{\mathrm{MAP}}$. Since $\boldsymbol{\theta}_{\mathrm{MAP}}$ is a maxima for the distribution $\mathbb{P}\left[\,\mathcal{D},\boldsymbol{\theta}\,\right]$, we can say $\bigtriangledown f(\boldsymbol{\theta}_{\mathrm{MAP}} = 0$. Therefore

$$\log\left(\mathbb{P}\left[\,\mathcal{D},\boldsymbol{\theta}\,\right]\right) \quad\approx\quad \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_{\mathrm{MAP}})^{\mathrm{T}} \bigtriangledown^2\, \log\left(\mathbb{P}\left[\,\mathcal{D},\boldsymbol{\theta}_{\mathrm{MAP}}\,\right]\right)(\boldsymbol{\theta}-\boldsymbol{\theta}_{\mathrm{MAP}})$$

Putting this back in the equation and simplifying, we get

$$\mathbb{P}\left[\,\boldsymbol{\theta}\,\middle|\,\mathcal{D}\,\right] \quad\approx\quad \frac{\exp\Big((\boldsymbol{\theta}-\boldsymbol{\theta}_{\mathrm{MAP}})^{\mathrm{T}}\left(-\bigtriangledown^2\,\log\left(\mathbb{P}\left[\,\mathcal{D},\boldsymbol{\theta}_{\mathrm{MAP}}\,\right]\right)\right)(\boldsymbol{\theta}-\boldsymbol{\theta}_{\mathrm{MAP}})\Big)}{\int\,\exp\Big((\boldsymbol{\theta}-\boldsymbol{\theta}_{\mathrm{MAP}})^{\mathrm{T}}\left(-\bigtriangledown^2\,\log\left(\mathbb{P}\left[\,\mathcal{D},\boldsymbol{\theta}_{\mathrm{MAP}}\,\right]\right)\right)(\boldsymbol{\theta}-\boldsymbol{\theta}_{\mathrm{MAP}})\Big)\,\mathrm{d}\boldsymbol{\theta}}$$

<++>

<++>