discussion4

Estimating Probabilities

- Bayes rule
- Maximum Likelihood Estimate (MLE)
- Maximum a Posterior (MAP)

logistics regression

Bayes rule

if binary. we can write as $D = a_i$ ⇐ 某种事件发生与否.

Given observations D, our goal is to estimate the parameter θ. Through Bayes rule, we have the following identity,

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

where we call $P(\theta)$ the prior, $P(\theta|D)$ the posterior and $P(D|\theta)$ the likelihood.

MLE

One approach to estimate probabilities is to maximize the likelihood as follows,

$$\hat{\theta}^{MLE} = \arg\max_{\theta} P(D|\theta),$$

which is the general definition of MLE.

Intuition

We observe training data D, we should choose the value of θ that makes D most probable.

An example

$X \sim Bern(p)$

$f(x \mid p) = \begin{cases} p^x (1-p)^{1-x} & x = 0, 1 \\ 0 & x \neq 0, 1 \end{cases}$

$D = \{x_1, \cdots x_n\}$ i.i.d.

$P(D \mid \theta) = \prod_{i=1}^{n} \theta^{x_i=1}(1-\theta)^{x_i=0}$

$= \theta^{\alpha_1}(1-\theta)^{\alpha_0}$

- X be a random variable for a coin, 1 or 0,
- θ is the probability of X taking 1, e.g., $P(X = 1) = \theta$, and unknown,
- D is the observations produced by flip a coin X $N = \alpha_1 + \alpha_0$ times where α_1 the number of $X = 1$,
- Assuming I.I.D.

Continuing

Likelihood is defined as $L(\theta) = P(D|\theta)$. With the conditions claimed before, we have the following formula,

$$L(\theta) = P(D|\theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}.$$

The MLE is to choose θ to maximize $P(D|\theta)$. For convenient, we take the log of $L(\theta)$,

$$l(\theta) = \ln L(\theta) = \alpha_1 \ln \theta + \alpha_0 \ln(1-\theta),$$

where $l(\theta)$ is called as log-likelihood. Since $l(\theta)$ is a concave function of θ, we just calculate the derivative of $l(\theta)$ with respect to θ,

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \frac{\partial \ln P(D|\boldsymbol{\theta})}{\partial \theta} \\ &= \frac{\partial \ln [\boldsymbol{\theta}^{\alpha_1}(1-\boldsymbol{\theta})^{\alpha_0}]}{\partial \theta}\end{aligned}$$

$$
\begin{aligned}
&= \frac{\partial \left[\alpha_1 \ln \theta + \alpha_0 \ln(1-\theta)\right]}{\partial \theta} \\
&= \alpha_1 \frac{\partial \ln \theta}{\partial \theta} + \alpha_0 \frac{\partial \ln(1-\theta)}{\partial \theta} \\
&= \alpha_1 \frac{\partial \ln \theta}{\partial \theta} + \alpha_0 \frac{\partial \ln(1-\theta)}{\partial (1-\theta)} \cdot \frac{\partial (1-\theta)}{\partial \theta} \\
\implies \frac{\partial \ell(\theta)}{\partial \theta} &= \alpha_1 \frac{1}{\theta} + \alpha_0 \frac{1}{(1-\theta)} \cdot (-1) \\
\implies \theta &= \frac{\alpha_1}{\alpha_1 + \alpha_0}
\end{aligned}
$$

Thus,

$$
\hat{\theta}^{MLE} = \arg\max_{\theta} P(D|\theta) = \arg\max_{\theta} \ln P(D|\theta) = \frac{\alpha_1}{\alpha_1 + \alpha_0}
$$

MAP

Given the observed data D and the prior $P(\theta)$, we want to maximize the posterior probability. By using Bayes rule, we have

$$\hat{\theta}^{MAP} = \arg\max_{\theta} P(\theta|D) = \arg\max_{\theta} \frac{P(D|\theta)P(\theta)}{P(D)}$$

Comparing the MLE algorithm, the only difference is the extra $P(\theta)$.

Intuition

Given new evidence, update the prior knowledge.

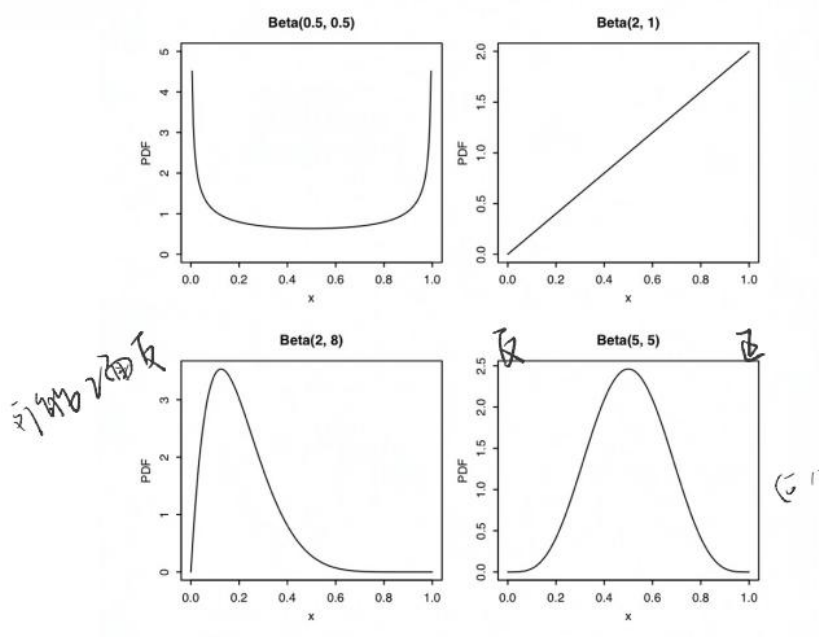
As in our coin flip example, the most common form of prior is a Beta distribution:

$$P(\theta) = \text{Beta}\left(\beta_0, \beta_1\right) = \frac{\theta^{\beta_1 - 1}(1-\theta)^{\beta_0 - 1}}{B\left(\beta_0, \beta_1\right)}.$$

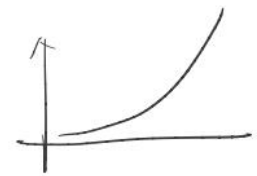
Definition 8.3.1 (Beta distribution)**.** An r.v. X is said to have the *Beta distribution* with parameters a and b, $a > 0$ and $b > 0$, if its PDF is

$$f(x) = \frac{1}{\beta(a,b)} x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1,$$

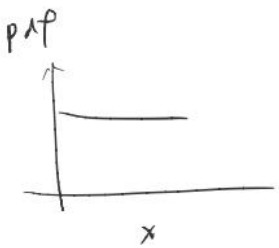
$$\beta(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx.$$



正反.
(3,0)
Beta(1,1) → Beta(4,1)



Beta(1,1)



共轭先验 beta分布经过贝叶斯后依旧是beta分布

$a + d_1 - 1$
$b + d_0 - 1$

Recall the expression for $P(D|\theta)$, we have:

$$\begin{aligned}
\hat{\theta}^{MAP} &= \arg\max_{\theta} P(D|\theta)P(\theta) \\
&= \arg\max_{\theta} \theta^{\alpha_1}(1-\theta)^{\alpha_0} \frac{\theta^{\beta_1-1}(1-\theta)^{\beta_0-1}}{B(\beta_0,\beta_1)} \\
&= \arg\max_{\theta} \frac{\theta^{\alpha_1+\beta_1-1}(1-\theta)^{\alpha_0+\beta_0-1}}{B(\beta_0,\beta_1)} \\
&= \arg\max_{\theta} \theta^{\alpha_1+\beta_1-1}(1-\theta)^{\alpha_0+\beta_0-1}.
\end{aligned}$$

Substitute $(\alpha_1+\beta_1-1)$ for α_1 and $(\alpha_0+\beta_0-1)$ for α_0 in $\hat{\theta}^{MLE}$, we have

$$\hat{\theta}^{MAP} = \arg\max_{\theta} P(D|\theta)P(\theta) = \frac{(\alpha_1+\beta_1-1)}{(\alpha_1+\beta_1-1)+(\alpha_0+\beta_0-1)}.$$

Shrinkage Methods – Discussion

Ridge and Lasso in the Bayes framework

$$\text{MLE: } \hat{\beta}^{MLE} = \text{argmax}_{\beta} \Pr(\mathbf{y}|\mathbf{X}, \beta) \leftarrow \text{Least squares}$$

$$\text{MAP: } \hat{\beta}^{MAP} = \text{argmax}_{\beta} \Pr(\mathbf{y}|\mathbf{X}, \beta)\Pr(\beta) \leftarrow \text{Ridge \& Lasso}$$

- Ridge regression
  - MAP with a prior $\Pr(\beta) = \mathcal{N}(\beta|0, \frac{1}{\lambda}\mathbf{I}_p)$ Gaussian distribution

$$\hat{\beta}^{ridge} = \text{argmax}_{\beta} \ln\big(\Pr(\mathbf{y}|\mathbf{X}, \beta)\Pr(\beta)\big)$$

$$= \text{argmax}_{\beta} \ln\left(\prod_{i=1}^{N} \mathcal{N}(y_i|x_i^T\beta, \sigma^2) \times \mathcal{N}(\beta|0, \frac{1}{\lambda}\mathbf{I}_p)\right)$$

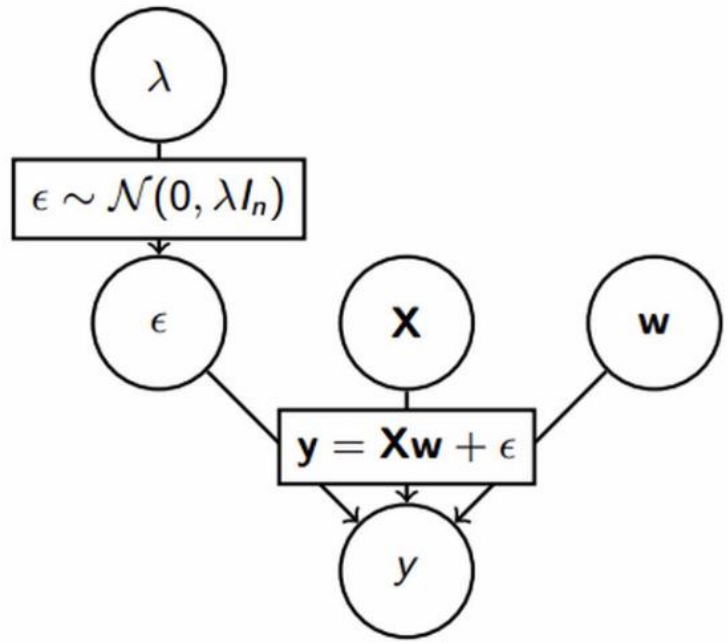
- Lasso
  - MAP with a prior $\Pr(\beta) = \frac{\lambda}{2} e^{-\lambda\|\beta\|_1}$ Laplacian distribution

$$\hat{\beta}^{lasso} = \text{argmax}_{\beta} \ln\left(\prod_{i=1}^{N} \mathcal{N}(y_i|x_i^T\beta, \sigma^2) \times \frac{\lambda}{2} e^{-\lambda\|\beta\|_1}\right)$$

Model

The regression model : $\mathbf{y} = \mathbf{X}\mathbf{w} + \epsilon$ can be see as the following model :

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \lambda) = \mathcal{N}(\mathbf{X}\mathbf{w}, \lambda) \text{ with } p(\epsilon) = \mathcal{N}(0, \lambda)$$



$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\delta} exp\left(-\frac{(\epsilon^{(i)})^2}{2\delta^2}\right)$$

$$\Rightarrow p(y^{(i)}|x^{(i)}; \theta) = \frac{1}{\sqrt{2\pi}\delta} exp\left(-\frac{(y^{(i)} - w^T x^{(i)})^2}{2\delta^2}\right)$$

$$
\begin{aligned}
L(w) &= p(\vec{y}|X;w) \\
&= \prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta) \\
&= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\delta} exp\left(-\frac{(y^{(i)}-w^T x^{(i)})^2}{2\delta^2}\right)
\end{aligned}
$$

$$
\begin{aligned}
l(w) &= \log L(w) \\
&= \log \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\delta} exp\left(-\frac{(y^{(i)}-w^T x^{(i)})}{2\delta^2}\right) \\
&= \sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi}\delta} exp\left(-\frac{(y^{(i)}-w^T x^{(i)})^2}{2\delta^2}\right) \\
&= m \log \frac{1}{\sqrt{2\pi}\delta} - \frac{1}{\delta^2} \cdot \frac{1}{2} \sum_{i=1}^{m} (y^{(i)}-w^T x^{(i)})^2
\end{aligned}
$$

$$
w_{MLE} = \arg\min_{w} \sum_{i=1}^{m} (y^{(i)}-w^T x^{(i)})^2
$$

Ridge--- MAP with Gaussian distribution

$$
\begin{aligned}
L(w) &= p(\vec{y}|X;w)p(w) \\
&= \prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta)p(w) \\
&= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\delta} exp\left(-\frac{(y^{(i)}-w^T x^{(i)})^2}{2\delta^2}\right) \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\alpha}} exp\left(-\frac{(w^{(j)})^2}{2\alpha}\right) \\
&= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\delta} exp\left(-\frac{(y^{(i)}-w^T x^{(i)})^2}{2\delta^2}\right) \frac{1}{\sqrt{2\pi\alpha}} exp\left(-\frac{w^T w}{2\alpha}\right)
\end{aligned}
$$

$$
\begin{aligned}
l(w) &= \log L(w) \\
&= m \log \frac{1}{\sqrt{2\pi}\delta} + n \log \frac{1}{\sqrt{2\pi}\alpha} - \frac{1}{\delta^2} \cdot \frac{1}{2} \sum_{i=1}^{m} (y^{(i)} - w^T x^{(i)})^2 - \frac{1}{\alpha} \cdot \frac{1}{2} w^T w \\
\Rightarrow w_{MAP_{Guassian}} &= \arg\min_{w} \left(\frac{1}{\delta^2} \cdot \frac{1}{2} \sum_{i=1}^{m} (y^{(i)} - w^T x^{(i)})^2 + \frac{1}{\alpha} \cdot \frac{1}{2} w^T w \right)
\end{aligned}
$$

Ridge $\frac{1}{n} \|y - w^T X\|_2 + \lambda \|w\|_2$

Lasso --- MAP with Laplace distribution

$$f(x \mid \mu, b) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$$

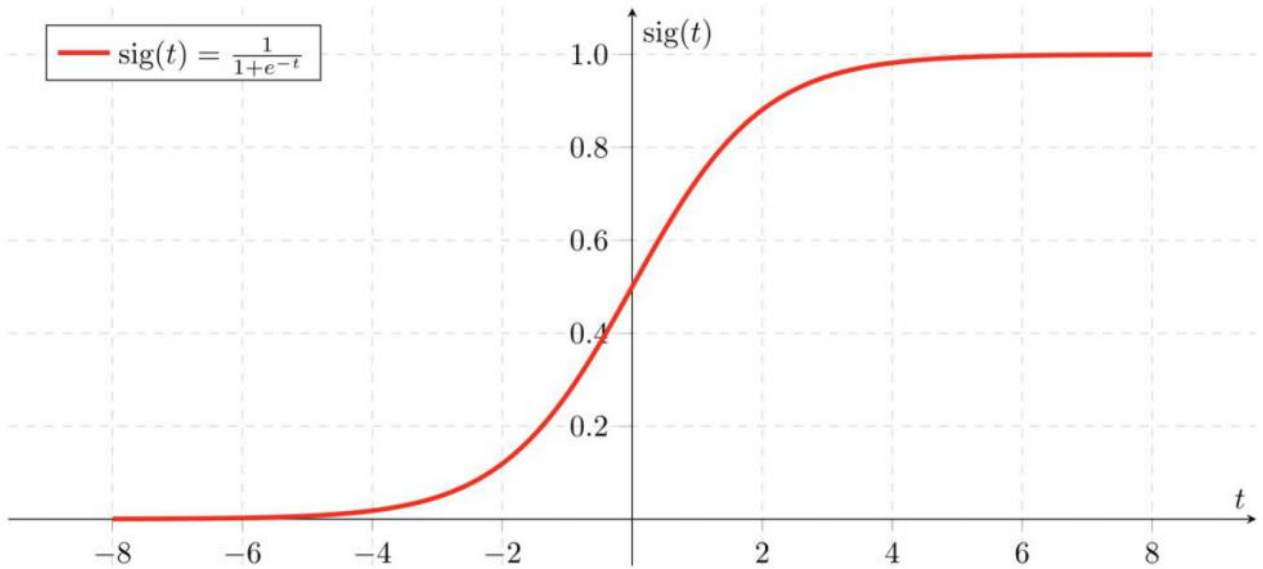
$$w_{MAP_{Laplace}} = \arg\min_{w}\left(\frac{1}{\delta^2}\cdot\frac{1}{2}\sum_{i=1}^{m}(y^{(i)} - w^T x^{(i)})^2 + \frac{1}{b^2}\cdot\frac{1}{2}\|w\|_1\right)$$

logistics regression

logit function

$$\text{map } \theta^T x(z) : -\infty \sim \infty \text{ to } 0 \sim 1$$

If ‘Z’ goes to infinity, Y(predicted) will become 1 and if ‘Z’ goes to negative infinity, Y(predicted) will become 0.



logistics regression

linear classification: find the hyperplane

$$\theta^T x + b = 0$$

threshold.

$\theta^T x \geq -b \rightarrow y = 0$

$\theta^T x < -b \rightarrow y = 1$

predict result $\theta^T x$

$$h_\theta(x) \Rightarrow g\left(\theta^T x\right) = \frac{1}{1+e^{-\theta^T x}} = \frac{1}{1+e^{-z}}$$

$$\begin{cases} p(Y=0|X) = \dfrac{1}{1+e^{-\theta^T x}} \\ p(Y=1|X) = \dfrac{e^{-\theta^T x}}{1+e^{-\theta^T x}} \end{cases} \rightarrow \text{probability distribution}$$

Binary Logistic Regression

We view each observasions y_i as an independent sample from a Bernoulli distribution $Y_i \sim Bern(p_i)$, (technically we mean $\hat{Y}_i|\mathbf{x}_i, \mathbf{w}$), where p_i is a function of $\mathbf{x}_i$.

We need a model for the dependency of p_i on $\mathbf{x}_i$. We have to enforce that p_i is a transformation of $\mathbf{x}_i$ that results in a number from 0 to 1 (ie. a valid probability). Hence p_i cannot be, say, linear in x_i. One way to do achieve the 0-1 normalization is by using the sigmoid function.

$$p_i = s(\mathbf{w}^T \mathbf{x}_i) = \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}_i}}$$

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} P(y_1, \cdots, y_n | \mathbf{x}_1, \cdots, \mathbf{x}_n, \mathbf{w})$$

$$= \arg\max_{\mathbf{w}} \prod_{i=1}^{n} P(y_i | \mathbf{x}_i, \mathbf{w}) \quad \text{iid.}$$

(max-log-likelihood).

$$= \arg\max_{\mathbf{w}} \ln[\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{(1-y_i)}]$$

$$= \arg\max_{\mathbf{w}} \sum_{i=1}^{n} y_i \ln p_i + (1-y_i)\ln(1-p_i)$$

$$= \arg\min_{\mathbf{w}} -\sum_{i=1}^{n} y_i \ln p_i + (1-y_i)\ln(1-p_i)$$

loss function.

Binary Logistic Regression

$$
\begin{aligned}
L(\mathbf{w}) &= -\sum_{i=1}^{n} y_i \ln p_i + (1 - y_i) \ln(1 - p_i) \\
\nabla_{\mathbf{w}} L(\mathbf{w}) &= \nabla_{\mathbf{w}} \left(-\sum_{i=1}^{n} y_i \ln p_i + (1 - y_i) \ln(1 - p_i) \right) \\
&= -\sum_{i=1}^{n} y_i \nabla_{\mathbf{w}} \ln p_i + (1 - y_i) \nabla_{\mathbf{w}} \ln(1 - p_i) \\
&= -\sum_{i=1}^{n} \frac{y_i}{p_i} \nabla_{\mathbf{w}} p_i - \frac{1 - y_i}{1 - p_i} \nabla_{\mathbf{w}} p_i
\end{aligned}
$$

Binary Logistic Regression

$$
\begin{aligned}
p_i &= s(\mathbf{w}^T x_i) = \frac{1}{1 + e^{-\mathbf{w}^T x_i}} \\
\nabla_z s(z) &= \nabla_z (1 + e^{-z})^{-1} \\
&= \frac{e^{-z}}{(1 + e^{-z})^2} \\
&= s(z)(1 - s(z)) \\
\nabla_{\mathbf{w}} L(\mathbf{w}) &= -\sum_{i=1}^{n} \frac{y_i}{p_i} \nabla_{\mathbf{w}} p_i - \frac{1 - y_i}{1 - p_i} \nabla_{\mathbf{w}} p_i \\
&= -\sum_{i=1}^{n} \left(\frac{y_i}{p_i} - \frac{1 - y_i}{1 - p_i} \right) p_i (1 - p_i) x_i \\
&= -\sum_{i=1}^{n} (y_i - p_i) x_i
\end{aligned}
$$