CS 189 Introduction to Machine Learning
Spring 2019 Jonathan Shewchuk

DIS3

1 Logistic posterior with exponential class conditionals

We have seen in class that Gaussian class conditionals lead to a logistic posterior that is linear or quadratic in X. Now, suppose the class conditionals are exponentially distributed with paramateres λ_i, i.e.

$$X|Y = i \sim \lambda_i \exp(-\lambda_i x), \quad \text{where } i \in \{0, 1\}$$
$$Y \sim \text{Bernoulli}(\pi)$$

Show that the posterior distribution of the class label given X is also a logistic function, however with a linear argument in X. Assuming 0-1 loss, what will the decision boundary look like (i.e., describe what the posterior probability plot looks like)?

Solution:

We are solving for $P(Y = 1|x)$. By Bayes Rule, we have

$$\begin{aligned} P(Y = 1|x) &= \frac{P(x|Y = 1)P(Y = 1)}{P(x|Y = 1)P(Y = 1) + P(x|Y = 0)P(Y = 0)} \\ &= \frac{1}{1 + \frac{P(Y=0)P(x|Y=0)}{P(Y=1)P(x|Y=1)}} \\ &= \frac{1}{1 + \frac{\lambda_0}{\lambda_1}\frac{1-\pi}{\pi}\exp\left(-\lambda_0 x + \lambda_1 x\right)} \end{aligned}$$

Looking at the bottom right equation, we have

$$\frac{\lambda_0}{\lambda_1}\frac{1-\pi}{\pi}\exp\left(-\lambda_0 x + \lambda_1 x\right) = \exp\left(-(\lambda_0 - \lambda_1)x + \log\left(\frac{\lambda_0}{\lambda_1}\frac{1-\pi}{\pi}\right)\right)$$

Now we see that we have a logistic function $\frac{1}{1+\exp\left(-h(x)\right)}$, where $h(x) = ax + b$ is linear (affine) in x. Since we are assuming 0-1 loss, we use the optimal classifier $f^*(x) = 1$ when $P(Y = 1|x) > P(Y = 0|x)$. Thus, the decision boundary can be found when $P(Y = 1|x) = P(Y = 0|x) = \frac{1}{2}$. This happens when $h(x) = 0$. Solving for x gives

$$\bar{x} = \frac{\log \frac{\lambda_0}{\lambda_1}\frac{1-\pi}{\pi}}{\lambda_0 - \lambda_1}.$$

If we assume $\lambda_0 > \lambda_1$, then the optimal classifier is

$$f^*(x) = \begin{cases} 1 & \text{if } x > \bar{x} \\ 0 & o.w. \end{cases}$$

2 Bayesian Decision Theory: Case Study

We want to design an automated fishing system that captures fish, classifies them, and sends them off to two different companies, Salmonites, Inc., and Seabass, Inc. For some reason we only ever catch salmon ($Y = 1$) and seabass ($Y = 2$). Salmonites, Inc. wants salmon, and Seabass, Inc. wants seabass. Given only the weights of the fish we catch, we want to figure out what type of fish it is using machine learning!

Let us assume that the weight of both seabass and salmon are both normally distributed (univariate Gaussian), given by the p.d.f.:

$$P(x|Y = i) = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}$$

We are given this data:

Data for salmon: $\{3, 4, 5, 6, 7\}$

Data for seabass: $\{5, 6, 7, 8, 9, 7 + \sqrt{2}, 7 - \sqrt{2}\}$

When we classify seabass incorrectly, it gets sent to Salmonites, Inc. who won't pay us for the wrong fish and sells it themselves. When we classify salmon incorrectly, it gets sent to SeaBass, Inc., who is nice and returns our fish. This situation gives rise to this loss matrix:

Predicted:

Truth:		**salmon**	**seabass**
	salmon	0	1
	seabass	2	0

(a) First, compute the sample mean $\hat{\mu}_i$ and variance $\hat{\sigma}_i^2$ for the univariate Gaussian in both the seabass and the salmon case. Also compute the empirical estimates of the priors $\hat{\pi}_i$.

$\hat{\mu}_1 =$
$\hat{\sigma}_1^2 =$
$\hat{\pi}_1 =$

$\hat{\mu}_2 =$
$\hat{\sigma}_2^2 =$
$\hat{\pi}_2 =$

(b) What is significant about $\hat{\sigma_1}$ and $\hat{\sigma_2}$?

(c) Next, find the decision rule when assuming a 0-1 loss function. Recall that a decision rule for the 0-1 loss function will minimize the probability of error.

(d) Now, find the decision rule using the loss matrix above. Recall that a decision rule, in general, minimizes the risk, or expected loss.

Solution:

(a) Sample mean $\hat{\mu}$:

$$\hat{\mu} = \frac{1}{N}\sum_i X_i$$

Sample covariance:

$$\hat{\sigma}^2 = \frac{1}{N}\sum_i (X_i - \hat{\mu})^2$$

Plugging in numbers for seabass and salmon: $\hat{\mu}_1 = 5$, $\hat{\mu}_2 = 7$, $\hat{\sigma}_1^2 = 2$, $\hat{\sigma}_2^2 = 2$

Calculating the priors: $\hat{\pi}_1 = 5/12$, $\hat{\pi}_2 = 7/12$

(b) They're the exact same, so a decision boundary between the two Gaussians characterized by them will be linear.

(c) Recall that assuming a 0-1 loss function results in choosing the class to minimize the probability of error, which means choosing according to this rule:

$$\text{If } \frac{p(Y=1|x)}{p(Y=2|x)} > 1, \text{ choose 1}$$

Because there is a linear decision boundary, we search for the value such that we classify everything to the right as seabass, and everything to the left as salmon. This boundary is the value of x such that $p(Y=1|x) = p(Y=2|x)$.

$$p(Y=1|x) = p(Y=2|x) \implies 5p(x|Y=1) = 7p(x|Y=2)$$

$$\frac{5}{\sqrt{2\pi}\sigma} exp\left(-\frac{1}{2}\frac{(x-5)^2}{\sigma^2}\right) = \frac{7}{\sqrt{2\pi}\sigma} exp\left(-\frac{1}{2}\frac{(x-7)^2}{\sigma^2}\right)$$

$$\ln(5) - \frac{1}{2\sigma^2}(x-5)^2 = \ln(7) - \frac{1}{2\sigma^2}(x-7)^2$$

$$4\ln\left(\frac{5}{7}\right) - x^2 + 10x - 25 = -x^2 + 14x - 49$$

$$4\ln\left(\frac{5}{7}\right) + 24 = 4x$$

$$x = \ln\left(\frac{5}{7}\right) + 6 \approx 5.66$$

The decision rule is: If $x > 5.66$, classify as Seabass! Otherwise classify as Salmon.

Note: Because we had the same variance for both class conditionals, the x^2 term canceled out. If that was not the case, then there would be 3 regions, and we would allocate 2 of them to one fish, 1 of them to the other, depending on the height of the posterior probabilities. A good exercise would be to try to draw this: two 1-D Gaussians with different variances.

(d) In the general case, we want to make the decision that minimizes risk. Thus, the decision boundary is located at where the risk of making either decision is equal, or:

$$R(\hat{y} = 1|x) = R(\hat{y} = 2|x)$$

Recall that $R(\hat{y} = i|x) = \sum_{j=1}^{C} \lambda_{ij} P(Y = j|x)$.

$$\lambda_{11} P(Y = 1|x) + \lambda_{12} P(Y = 2|x) = \lambda_{21} P(Y = 1|x) + \lambda_{22} P(Y = 2|x)$$

$$2 \cdot P(Y = 2|x) = 1 \cdot P(Y = 1|x)$$

$$2 \cdot \frac{7}{12}\mathcal{N}(7, 2) = 1 \cdot \frac{5}{12}\mathcal{N}(5, 2)$$

Solving this like part b), we get that $x = 6 + \ln\left(\frac{5}{14}\right) \approx 4.97$. Thus, if the weight is greater than 4.97, we classify it as seabass and if not, we classify it as salmon.

3 MLE vs. MAP

Maximum Likelihood Estimation (MLE) and Maximum A Posteriori (MAP) are both methods to estimate some parameters in a probablistic settings. MLE is a frequentist approach that draws conclusions from sample data by emphasizing the frequency or proportion of the data. Let the data be X and the parameter be θ, MLE assumes a likelihood distribution $P(X|\theta)$ and tries to maximize it:

$$\begin{aligned}\theta_{\text{MLE}} &= \arg\max_{\theta} \ P(X|\theta) \\ &= \arg\max_{\theta} \prod_{i=1}^{n} P(x_i|\theta) \\ &= \arg\max_{\theta} \sum_{i=1}^{n} \log P(x_i|\theta)\end{aligned}$$

where the last line follows that taking logarithm does not change the result of a maximization and usually summation is easier to handle than products. In fact, taking sample mean and sample variance for a data cluster is taking a MLE approach.

On the other hand, MAP takes a more Baynesian approach, an alternate approach to frequentist inference. MAP assumes that the parameter θ is also a random variable and has its own distribution and an underlying posterior distribution $P(\theta|X)$ with the data. Recall that using Bayes' rule, the posterior distribution can be seen as the product of likelihood and prior:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \propto \underbrace{P(X|\theta)}_{\text{likelihood}} \underbrace{P(\theta)}_{\text{prior}}$$

MAP tries to infer the parameter by maximizing the posterior distribution:

$$\begin{aligned}\theta_{\text{MAP}} &= \arg\max_{\theta} \ P(\theta|X) \\ &= \arg\max_{\theta} \ P(X|\theta)P(\theta) \\ &= \arg\max_{\theta} \prod_{i=1}^{n} P(x_i|\theta)P(\theta) \\ &= \arg\max_{\theta} \sum_{i=1}^{n} \log P(x_i|\theta) + n \log P(\theta)\end{aligned}$$

Note that since both of these methods are point estimates (they yield a value rather than a distribution), neither of them are completely Bayesian. A faithful Bayesian would use a model that yields a posterior distribution over all possible values of θ, instead of maximizing the function.

Now suppose we have a coin with unknown bias θ. We are trying to find the bias of the coin by maximizing the underlying distribution. You tossed the coin $n = 10$ times and 3 of the tosses came as heads.

(a) What is your estimation of the bias of the coin, θ, if you are using MLE?

(b) Suppose we know that the bias of the coin is $\theta \sim \mathcal{N}(0.8, 0.3)$, i.e., we are rather sure that the bias should be around 0.8. Now what is your estimation of θ if you are using MAP model? You can leave your result as a polynomial equation on θ.

(c) What if our prior is $\theta \sim \mathcal{N}(0.5, 0.3)$ or $\mathcal{N}(0.8, 1)$. How does the difference between MAP and MLE change and why?

(d) What if our prior is that θ is uniformly distributed in the range $(0, 1)$? How would you view MLE from a more Bayeisan perspective?

Solution:

(a)

$$P(x|\theta) = \theta^x(1-\theta)^{(n-x)} = \theta^3(1-\theta)^7.$$

Taking the logarithm for easier computation, we have

$$\ln P(x|\theta) = 3\ln\theta + 7\ln(1-\theta).$$

This is a concave function and thus the maximum is achieved by setting the derivative w.r.t. θ to 0:

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\ln P(x|\theta) = \frac{3}{\theta} - \frac{7}{1-\theta} = 0.$$

Therefore,

$$\hat{\theta}_{\text{MLE}} = 0.3.$$

(b) Now take into account the prior distribution:

$$\begin{aligned} P(\theta|x) &\propto P(x|\theta)P(\theta) \\ &= \theta^x(1-\theta)^{n-x}\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2}\right] \\ &= \theta^3(1-\theta)^7\frac{1}{\sqrt{2\pi}0.3}\exp\left[-\frac{(\theta-0.8)^2}{2\times 0.09}\right]. \end{aligned}$$

Taking the logarithm,

$$\ln P(x,\theta) = 3\ln\theta + 7\ln(1-\theta) - \frac{1}{2}\ln(2\pi\times 0.3) - \frac{(\theta-0.8)^2}{2\times 0.09}.$$

Taking the derivative w.r.t. θ,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\ln P(x|\theta) = \frac{3}{\theta} - \frac{7}{1-\theta} - \frac{\theta-0.8}{0.09} = 0.$$

Solving the equation yields

$$\hat{\theta}_{\text{MAP}} \approx 0.406.$$

$\hat{\theta}$ is now larger because we are assuming a larger prior.

(c) $\hat{\theta}_{\text{MAP}} \approx 0.340$ for $\mathcal{N}(0.5, 0.3)$ and ≈ 0.31 for $\mathcal{N}(0.8, 1)$. For $\mathcal{N}(0.5, 0.3)$, the prior is less distant from the experiment result; for $\mathcal{N}(0.8, 1)$, the prior is weaker due to a larger variance. Therefore, the difference between the two models will decrease. In fact, in many real world scenarios, we don't have a prior or the prior is not very strong. In those cases, the difference between the two models is not as large as shown in this problem.

(d) The two results will be the same since the prior terms $P(\theta)$ is uniform and can be canceled out. From a Bayesian perspective, the result of MLE can be seen as a special case of MAP with a uniform prior.