CS 229, Spring 2022
Section #2 Solutions: Generalized Linear Models and Gaussian Discriminant Analysis

---

1. **Generalized Linear Models**

   In lecture, we have seen that many of the distributions that we commonly use to model the world, such as Gaussian, Bernoulli, Exponential, and Beta distributions, are all part of the **Exponential Family** of distributions. These distributions can be written in the general form:

   $$P(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$$

   We have also seen the idea of **Generalized Linear Models**, a broad family of models that utilize the notion of exponential family distributions.

   (a) What are the three assumptions/design choices we make when constructing a generalized linear model about the conditional distribution of y given x and our model?

   Answer:

   i. $y \mid x;\theta \sim \text{ExponentialFamily}(\eta)$. I.e., given x and θ, the distribution of y follows some exponential family distribution, with parameter η.

   ii. The natural parameter η and the inputs x are related linearly: $\eta = \theta^T x$. (Or, if η is vector-valued, then $\eta_i = \theta_i^T x$)

   iii. At test time, we make a prediction $h(x)$ as the expected value of $T(y)$ given x. In most of our examples, we will have $T(y) = y$, so this means we predict $h(x) = E[y|x]$.

   (b) You would like to create a generalized linear model such that y conditioned on x is modeled as an exponential distribution, i.e.

   $$y \mid x;\theta \sim \text{Exponential}(\lambda)$$

   The probability density function for an exponential random variable is defined to be

   $$p(y;\lambda) = \begin{cases} \lambda e^{-\lambda y} & y \geq 0 \\ 0 & y < 0 \end{cases}$$

   for $\lambda > 0$.

   i. Show that the exponential distribution is a member of the exponential family by writing the PDF in the exponential family format.

   ii. Confirm that the gradient of the log likelihood function with respect to the parameter θ has the same form as other exponential family distributions we have seen in class: $\frac{\partial}{\partial \theta_j}\ell(\theta) = \sum_{i=1}^{m}\left(y^{(i)} - h_\theta(x^{(i)})\right)x_j$

   Hint: The expectation of an exponential distribution is $1/\lambda$

Answer:

i. For $y > 0$,

$$\begin{aligned} p(y;\lambda) &= \lambda e^{-\lambda y} \\ &= e^{\log\lambda} e^{-\lambda y} \\ &= e^{(-\lambda)y-(-\log(\lambda)} \end{aligned}$$

We see that $\eta = -\lambda$, $T(y) = y$, $a(\eta) = -\log(\lambda) = -\log(-\eta)$, and $b(y) = 1$.

ii.

$$\begin{aligned} \ell(\theta) &= \log\ p(\vec{y} \mid X;\theta) \\ &= \log \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)};\theta) \\ &= \log \prod_{i=1}^{m} e^{(-\lambda)y^{(i)}+\log(\lambda)} \\ &= \log \prod_{i=1}^{m} e^{(\theta^T x^{(i)})y^{(i)}+\log(-\theta^T x^{(i)})} \\ &= \sum_{i=1}^{m} (\theta^T x^{(i)})y^{(i)} + \log(-\theta^T x^{(i)}) \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial\theta_j}\ell(\theta) &= \sum_{i=1}^{m} {x^{(i)}}_j y^{(i)} - \frac{1}{(-\theta^T x^{(i)})} {x^{(i)}}_j \\ &= \sum_{i=1}^{m} (y^{(i)} - \frac{1}{(-\theta^T x^{(i)})}) {x^{(i)}}_j \\ &= \sum_{i=1}^{m} \left(y^{(i)} - h_\theta(x^{(i)})\right) {x^{(i)}}_j \end{aligned}$$

From i. we see that $\eta = -\lambda$. By design choice 3 iii from part a, $\eta = \theta^T x$, so $\lambda = -\theta^T x$. By design choice 2, our $h(x) = E[y|x] = \frac{1}{\lambda} = \frac{1}{-\theta^T x}$.

2. **Generative Models (GDA)**

Generative learning algorithms model the full joint distribution $p(x, y)$ by modeling $p(y)$ and $p(x|y)$ for labels y and data x. This may be contrasted with **discriminative learning algorithms**, which directly model $p(y|x)$. Advantages of generative models include the ability to generate new data once the model is trained and potential improvements in performance and data efficiency (the amount of training data required to learn "well") if our modeling assumptions are accurate (or approximately accurate). A potential disadvantage of generative models is that performance may suffer if our assumptions are inaccurate.

(a) **Gaussian Discriminant Analysis (GDA):** In GDA, we make a **STRONG** assumption that the data came from 2 Gaussian Distributions with equal covariance.[1] That is:

$$\begin{aligned} p(y) &= \phi^y(1-\phi)^{(1-y)} \\ p(\vec{x}|y=0) &= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu}_0)^T\Sigma^{-1}(\vec{x}-\vec{\mu}_0)\right) \\ p(\vec{x}|y=1) &= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu}_1)^T\Sigma^{-1}(\vec{x}-\vec{\mu}_1)\right) \end{aligned}$$

i. Consider the 1-dimensional case. We have $\Sigma = [\sigma^2]$. Show that the posterior distribution is of the form

$$p(y=1|x;\theta) = \frac{1}{1+\exp(-(\theta^T x+\theta_0))}$$

This implies that the decision boundary is linear because if $p(y=1|x;\theta) = p(y=0|x)$ then $p(y=1|x;\theta) = \frac{1}{1+\exp(-(\theta^T x+\theta_0))} = .5$ and $\theta^T x + \theta_0 = 0$.

[1] It is also possible to model these distributions to have difference covariance matrices, though it is more complicated.

Answer: First notice that (for convenience, drop the θ from $p(y = 1 \mid x; \theta)$)

$$\begin{aligned}
p(y=1|x) &= \frac{p(y=1,x)}{p(x)} \\
&= \frac{p(x|y=1)p(y=1)}{p(x|y=1)p(y=1)+p(x|y=0)p(y=0)} \\
&= \frac{\phi \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2}\frac{(x-\mu_1)^2}{\sigma^2})}{\phi \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2}\frac{(x-\mu_1)^2}{\sigma^2})+(1-\phi)\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2}\frac{(x-\mu_0)^2}{\sigma^2})} \\
&= \frac{1}{1+\frac{(1-\phi)}{\phi}\exp(-\frac{1}{2}\frac{(x-\mu_0)^2}{\sigma^2}+\frac{1}{2}\frac{(x-\mu_1)^2}{\sigma^2})} \\
&= (1+\exp(\log(\frac{1-\phi}{\phi})-\frac{1}{2}\frac{(x-\mu_0)^2-(x-\mu_1)^2}{\sigma^2}))^{-1} \\
&= (1+\exp(\log(\frac{1-\phi}{\phi})-\frac{1}{2}\frac{x^2-2\mu_0 x+\mu_0^2-x^2+2\mu_1 x+\mu_1^2}{\sigma^2}))^{-1} \\
&= (1+\exp(\log(\frac{1-\phi}{\phi})-\frac{1}{2}\frac{-(2\mu_0-2\mu_1)x+(\mu_0^2-\mu_1^2)}{\sigma^2}))^{-1} \\
&= (1+\exp(-(-\log(\frac{1-\phi}{\phi})-\frac{\mu_0-\mu_1}{\sigma^2}x+\frac{(\mu_0^2-\mu_1^2)}{2\sigma^2})))^{-1}
\end{aligned}$$

Thus we can see that

$$\begin{aligned}
\theta &= -\frac{\mu_0-\mu_1}{\sigma^2} \\
\theta_0 &= \frac{(\mu_0^2-\mu_1^2)}{2\sigma^2}-\log(\frac{1-\phi}{\phi})
\end{aligned}$$

So we have a linear decision boundary. Why is it that if we can write the posterior distribution in that form, we can see that the decision boundary is linear? Because we can set that probability to 0.5 and show that $\theta^T x + \theta_0 = 0$ is a linear equation for the decision boundary (set of x where the predicted probability is 0.5).

ii. (PSET question setup) Show that the decision boundary is linear in a multi-dimensional case.

Answer: We will not show the solution to this question because it is a homework question. But the set up is similar to the previous problem

iii. (Midterm Fall 2017) Consider a simpler case where we have a single feature for each example but we assume unequal covariances within each label, that is, $\Sigma_0 = [\sigma_0^2]$ and $\Sigma_1 = [\sigma_1^2]$. Show that the decision boundary is quadratic.

Answer: To find the decision boundary, we simply set $p(y = 1|x; \eta_0, \eta_1) = 1/2$ and

solve. Doing so gives the following:

$$\begin{aligned}
p(y=1|x) &= \frac{p(x|y=1)p(y=1)}{p(x|y=1)p(y=1)+p(x|y=0)p(y=0)} \\
&= \frac{\phi \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(\frac{-(x-\mu_1)^2}{2\sigma_1^2}\right)}{\phi \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(\frac{-(x-\mu_1)^2}{2\sigma_1^2}\right)+(1-\phi)\frac{1}{\sqrt{2\pi}\sigma_0}\exp\left(\frac{-(x-\mu_0)^2}{2\sigma_0^2}\right)} \\
&= \frac{1}{1+\frac{\sigma_1(1-\phi)}{\sigma_0\phi}\exp\left(\frac{(x-\mu_1)^2}{2\sigma_1^2}-\frac{(x-\mu_0)^2}{2\sigma_0^2}\right)} \\
&= \left[1+\exp\left(\ln\left(\frac{\sigma_1(1-\phi)}{\sigma_0\phi}\right)\right)\exp\left(\frac{(x-\mu_1)^2}{2\sigma_1^2}-\frac{(x-\mu_0)^2}{2\sigma_0^2}\right)\right]^{-1} \\
&= \left[1+\exp\left(\ln\left(\frac{\sigma_1(1-\phi)}{\sigma_0\phi}\right)+\frac{(x-\mu_1)^2}{2\sigma_1^2}-\frac{(x-\mu_0)^2}{2\sigma_0^2}\right)\right]^{-1} = 1/2
\end{aligned}$$

We can now rearrange the equation and get the following:

$$\begin{aligned}
\left[1+\exp\left(\ln\left(\frac{\sigma_1}{\sigma_0}\frac{(1-\phi)}{\phi}\right)+\frac{(x-\mu_1)^2}{2\sigma_1^2}-\frac{(x-\mu_0)^2}{2\sigma_0^2}\right)\right] = 2 \Rightarrow \\
\exp\left(\ln\left(\frac{\sigma_1}{\sigma_0}\frac{(1-\phi)}{\phi}\right)+\frac{(x-\mu_1)^2}{2\sigma_1^2}-\frac{(x-\mu_0)^2}{2\sigma_0^2}\right) = 1 \Rightarrow \\
\ln\left(\frac{\sigma_1}{\sigma_0}\frac{(1-\phi)}{\phi}\right)+\frac{(x-\mu_1)^2}{2\sigma_1^2}-\frac{(x-\mu_0)^2}{2\sigma_0^2} = 0 \Rightarrow \\
\left(\frac{1}{2\sigma_1^2}-\frac{1}{2\sigma_0^2}\right)x^2+\left(\frac{\mu_0}{\sigma_0^2}-\frac{\mu_1}{\sigma_1^2}\right)x+\left(\frac{\mu_1^2}{2\sigma_1^2}-\frac{\mu_0^2}{\sigma_0^2}+\ln(\frac{\sigma_1}{\sigma_0}\frac{(1-\phi)}{\phi})\right) = 0
\end{aligned}$$

which is a quadratic.