Discussion 9
Probabilistic Machine Learning, Spring 2018

1 Jensen's Inequality

Let Y be a positive random variable and let $p > q \geq 1$. Relate $\mathbb{E}\left[Y^p\right]^{1/p}$ to $\mathbb{E}\left[Y^q\right]^{1/q}$ by an inequality.

Solution: $f(Y) := Y^{p/q}$ is convex since the second derivative is positive: $\frac{d^2}{dY^2} f(Y) = \frac{p}{q}(\frac{p}{q} - 1)Y^{p/q-2} > 0$ since $p > q \geq 1$. By Jensen's inequality we have:

$$\begin{aligned}
\mathbb{E}\left[f(Y^q)\right] &\geq f(\mathbb{E}\left[Y^q\right]) \\
\mathbb{E}\left[(Y^q)^{p/q}\right] &\geq \mathbb{E}\left[Y^q\right]^{p/q} \\
\mathbb{E}\left[Y^p\right] &\geq \mathbb{E}\left[Y^q\right]^{p/q} \\
\mathbb{E}\left[Y^p\right]^{1/p} &\geq \mathbb{E}\left[Y^q\right]^{1/q}
\end{aligned}$$

2 Gaussian Mixture Models

Assume data is generated by two univariate Gaussian distributions, the first with mean 0 and variance 1, the second with mean 0 and variance 1/2. Let w denote the mixing weight. If there were a single observation x_1, what is the likelihood function and the maximum likelihood estimate $\hat{w}$ of w?

Solution: The likelihood is:

$$\begin{aligned}
L(w \mid x_1) = p(X_1 = x_1 \mid w) &= \frac{w}{\sqrt{2\pi}} e^{-\frac{1}{2}x_1^2} + \frac{1-w}{\sqrt{\pi}} e^{-x_1^2} \\
&= \left[\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_1^2} - \frac{1}{\sqrt{\pi}} e^{-x_1^2}\right] w + \frac{1}{\sqrt{\pi}} e^{-x_1^2}.
\end{aligned}$$

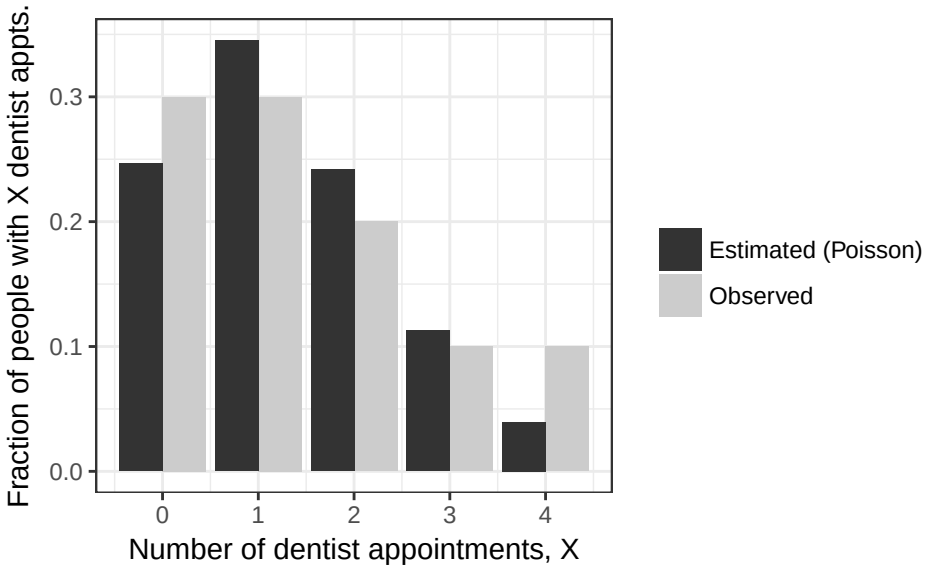
This is the equation of a line. If the slope is positive, then w should be as large as possible. If the slope is negative, then w should be as small as possible. One can show the slope will be positive if $x_1^2 > \log 2$. Therefore, since $w \in [0, 1]$, the solution is:

$$\hat{w} = \begin{cases} 1 & x_1^2 > \log 2 \\ 0 & \text{else} \end{cases} \tag{1}$$

3 Expectation Maximization (EM) for a Mixture of Two Different Distributions

Suppose $N = 100$ people were surveyed about how many dentist appointments they made in the past year. One way to model this data to assume the responses $\{x_i\}_{i=1}^{100}$ are drawn i.i.d. from a Poisson distribution, so $p(X_i = x_i \mid \lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$. Figure 1 shows a histogram of the observed number appointments and the predicted number of appointments from maximizing the assumed likelihood.

Figure 1



(a) Derive the maximum likelihood estimator $\hat{\lambda}$ for λ.

Solution: The log likelihood is given by:

$$\begin{aligned}
\log p(x_1, \ldots, x_N \mid \lambda) &= \log \prod_{i=1}^{N} p(x_i \mid \lambda) \\
&= \sum_{i=1}^{N} \log p(x_i \mid \lambda) \\
&= \sum_{i=1}^{N} -\lambda + x_i \log(\lambda) - \log(x_i!) \\
&= -N\lambda + \log(\lambda) \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \log(x_i!)
\end{aligned}$$

Taking the derivative with respect to λ and setting the result to zero we have:

$$\begin{aligned}
-N + \frac{1}{\hat{\lambda}} \sum_{i=1}^{N} x_i &= 0 \\
\hat{\lambda} &= \frac{1}{N} \sum_{i=1}^{N} x_i
\end{aligned}$$

So, $\hat{\lambda}$ is the sample mean.

We will now try a new model. Suppose that each person i is one of two types, denoted by a latent variable Z_i:

- If $Z_i = 1$, then person i never goes to the dentist (with probability 1), so $p(X_i = x_i \mid Z_i = 1, \lambda) = \mathbb{1}_{[x_i=0]}$.
- If $Z_i = 2$, then $p(X_i = x_i \mid Z_i = 2, \lambda) = \text{Poisson}(\lambda)$, as before.

We model each person as a mixture of these two types, letting $w = p(Z_i = 1 \mid \lambda)$ and $1 - w = p(Z_i = 2 \mid \lambda)$ denote the mixture weights. In general, the presence of a latent variable like Z_i can make it difficult to maximize the likelihood. We use the EM algorithm as a remedy to this problem.

E-step

In this step we compute the probability of each type assignment for each person. That is, we compute $\gamma_{i,k} := p(Z_i = k \mid X_i = x_i, \lambda)$ for all people $i \in \{1, \ldots, N\}$ and all types $k \in \{1, 2\}$.

(c) Write a formula for $p(X_i = x_i \mid Z_i = k, \lambda)$, the likelihood of observing outcome x_i given that person i is of type $Z_i = k$.

Solution:

$$p(X_i = x_i \mid Z_i = k, \lambda) = \begin{cases} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} & Z_i = 1 \\ \mathbb{1}_{[x_i=0]} & Z_i = 2 \end{cases}$$

(d) Write a formula for $p(X_i = x_i \mid \lambda)$, the likelihood of observing outcome x_i.

Solution:

$$\begin{aligned} p(X_i = x_i \mid \lambda) &= \sum_{k=1}^{2} p(X_i = x_i \mid Z_i = k, \lambda) p(Z_i = k \mid \lambda) \\ &= w\mathbb{1}_{[x_i=0]} + (1 - w)\frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \end{aligned}$$

(e) Write a formula for the type assignments $\gamma_{i,k} := p(Z_i = k \mid X_i = x_i, \lambda)$.

Solution:

$$\begin{aligned} \gamma_{i,k} :&= p(Z_i = k \mid X_i = x_i \lambda) \\ &= \frac{p(X_i = x_i \mid Z_i = k, \lambda) p(Z_i = k \mid \lambda)}{p(X_i = x_i \mid \lambda)} \\ &= \begin{cases} \frac{w\mathbb{1}_{[x_i=0]}}{w\mathbb{1}_{[x_i=0]} + (1-w)\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}} & Z_i = 1 \\ \frac{(1-w)\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}}{w\mathbb{1}_{[x_i=0]} + (1-w)\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}} & Z_i = 2 \end{cases} \end{aligned}$$

M-step

In this step we maximize a lower bound of the log likelihood:

$$A(w, \lambda) = \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_{i,k} \log \frac{p(X_i = x_i, Z_i = k \mid \lambda)}{\gamma_{i,k}}$$

over w and λ. Note that in this step the type assignments $\gamma_{i,k}$ are fixed.

(f) Maximize $A(w, \lambda)$ in λ. How does this compare to the maximum likelihood estimate when there was no latent variable (derived in part (a))?

Solution: Notice we can write:

$$\arg\max A(w,\lambda) = \arg\max \sum_{i=1}^{N}\sum_{k=1}^{2} \gamma_{i,k} \log \frac{p(X_i = x_i, Z_i = k \mid \lambda)}{\gamma_{i,k}}$$

$$= \arg\max \sum_{i=1}^{N}\sum_{k=1}^{2} \gamma_{i,k} \log p(X_i = x_i, Z_i = k \mid \lambda)$$

$$= \arg\max \sum_{i=1}^{N}\sum_{k=1}^{2} \gamma_{i,k} \log p(X_i = x_i \mid Z_i = k, \lambda) p(Z_i = k \mid \lambda)$$

$$= \arg\max \sum_{i=1}^{N} \gamma_{i,k=1} \log p(X_i = x_i \mid Z_i = 1, \lambda) p(Z_i = 1 \mid \lambda) + \gamma_{i,k=2} \log p(X_i = x_i \mid Z_i = 2, \lambda) p(Z_i = 2 \mid \lambda)$$

$$= \arg\max \sum_{i=1}^{N} \gamma_{i,k=1} \log \left(w \mathbb{1}_{[x_i=0]}\right) + \gamma_{i,k=2} \log \left((1-w)\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right)$$

$$= \arg\max_{\lambda} \sum_{i=1}^{N} \gamma_{i,k=1} \log \left(w \mathbb{1}_{[x_i=0]}\right) + \gamma_{i,k=2} \left(\log(1-w) - \lambda + x_i \log(\lambda) + \log(x_i!)\right) \tag{2}$$

Taking the derivative and setting the result to zero:

$$\frac{\partial A(w,\lambda)}{\partial \lambda} = \sum_{i=1}^{N} \gamma_{i,k=2} \left(-1 + x_i/\hat{\lambda}\right) = 0$$

$$\hat{\lambda} = \frac{\sum_{i=1}^{N} x_i \gamma_{i,k=2}}{\sum_{i=1}^{N} \gamma_{i,k=2}}$$

If $\gamma_{i,k=2}$ were always 1 (i.e., if each person were of type 2 with probability 1), then we would have the sample mean as in the maximum likelihood case without a latent variable.

(g) Maximize $A(w,\lambda)$ in w. How does this compare to the mixture of Gaussian distributions case, as derived in class?

Solution: Proceeding from Equation (2), we can drop terms not related to w:

$$\arg\max_{w} A(w,\lambda) = \arg\max_{w} \sum_{i=1}^{N} \gamma_{i,k=1} \log(w) + \gamma_{i,k=2} \log(1-w)$$

Taking the derivative and setting the result to zero:

$$\frac{\partial A(w,\lambda)}{\partial w} = \sum_{i=1}^{N} \frac{\gamma_{i,k=1}}{\hat{w}} - \frac{\gamma_{i,k=2}}{1-\hat{w}} = 0$$

$$\hat{w} = \frac{1}{N}\sum_{i=1}^{N} \gamma_{i,k=1}$$

Notice this is the same as in the mixture of Gaussians case.

4 Basics of Neural Networks

4.1

- A perceptron is guaranteed to perfectly learn a given linearly seperable function within a finite number of traning steps.

- For effective training of a neral network, the network should have at least 5-10 times as many weights as there are training samples.
- A single perceptron can coumpute the XOR function.
- The more hidden-layer units a BPN (BackPropagation Neural Network) has, the better it can predict desired outputs for new inputs that it was not trained with.
- In backpropagation learning, we should start with a small learning parameter η and slowly increase it during the learning process.
- A three-layer BPN with 5 neurons in each layer has a total of 50 connections and 50 weights.
- The backpropagation learning algorithm is based on the gradient-descent method.
- Some conflicts among training exemplars in a BPN can be resolved by adding features to the input vectors and adding input layer neurons to the network.

Solution:
T F F F F T T T

4.2

Derive the derivative of the tanh activation function

$$f(x) = \frac{2}{1 + e^{-x}} - 1$$

Can it be expressed as a function of $f(x)$? Explain.

Solution: Let

$$g(x) = \frac{2}{1 + e^{-x}}$$

$$f(x) = 2g(x) - 1 \rightarrow f'(x) = 2g'(x) = 2g(x)(1 - g(x))$$

$$g = 0.5(f + 1) \rightarrow f'(x) = 0.5(1 - f^2)$$