CS 189: Introduction to Machine Learning - Discussion 5

1. Review: Multivariate Gaussian

   (a) **True/False** If $X_1$ and $X_2$ are both normally distributed and independent, then $(X_1, X_2)$ must have multivariate normal distribution.

   **Solution:** True.
   Since $X_1$ and $X_2$ are independent with each other, so we have

   $$\begin{aligned} P(X_1, X_2) &= P(X_1)P(X_2) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(X_1-\mu_1)^2}{2\sigma_1^2}} \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(X_2-\mu_2)^2}{2\sigma_2^2}} \\ &\sim \mathcal{N}\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{pmatrix}\right) \end{aligned}$$

   (b) **True/False** If $(X_1, X_2)$ has multivariate normal distribution, then $X_1$ and $X_2$ are independent.

   **Solution:** False. If the off diagonal elements of the covariance matrix $\Sigma$ are not zeros, it means $cov(X_1, X_2) \neq 0$. Then they are not independent.

   (c) **Affine Transformations** Suppose $\mathbf{X} = [X_1, X_2, \cdots, X_n]^T$ is a n-dimensional random vector which has multivariate Gaussian distribution. Given $\mathbf{X} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$ and $\mathbf{y} = \mathbf{c} + \mathbf{BX}$ is an affine transformation of $\mathbf{X}$, where c is an $M \times 1$ vector of constants and $\mathbf{B}$ is a constant $M \times N$ matrix, what is the expectation and variance of $\mathbf{y}$?

   **Solution:**

   $$\begin{pmatrix} y_1 \\ \vdots \\ y_M \end{pmatrix} = \begin{pmatrix} c_1 \\ \vdots \\ c_M \end{pmatrix} + \begin{pmatrix} \mathbf{B_1} \\ \vdots \\ \mathbf{B_M} \end{pmatrix} \begin{pmatrix} X_1 \\ \vdots \\ X_N \end{pmatrix} = \begin{pmatrix} c_1 \\ \vdots \\ c_M \end{pmatrix} + \begin{pmatrix} \mathbf{B_1 X} \\ \vdots \\ \mathbf{B_M X} \end{pmatrix} = \begin{pmatrix} c_1 + \mathbf{B_1 X} \\ \vdots \\ c_M + \mathbf{B_M X} \end{pmatrix}$$

   So, $E\mathbf{y_i} = E(c_i + \mathbf{B_i X}) = c_i + E(\mathbf{B_i X})$ where $\mathbf{B_i}$ is 1xN vector and $\mathbf{X}$ is Nx1 vector and$\mathbf{B_i X} = \sum_{j=1}^{N} B_{ij} X_j$ for $i = \{1, \cdots, M\}$.
   $E(\mathbf{B_i X}) = \sum_{j=1}^{N} B_{ij} E(X_j) = \mathbf{B_i} E(\mathbf{X}) = \mathbf{B_i}\mu$
   So, we have $E(y_i) = c_i + \mathbf{B_i}\mu$. Therefore, $E(\mathbf{y}) = \mathbf{c} + \mathbf{B}\mu$

Similarly,

$$\begin{aligned} Var(\mathbf{y}) &= E[(\mathbf{Y}-\mathbf{EY})(\mathbf{Y}-\mathbf{EY})^{\mathbf{T}}] = E[(\mathbf{BX}-\mathbf{B}\mu)(\mathbf{BX}-\mathbf{B}\mu)^{\mathbf{T}}] \\ &= E(\mathbf{B}(\mathbf{X}-\mu)(\mathbf{X}-\mu)^{\mathbf{T}}\mathbf{B}) = \mathbf{B}E[(\mathbf{X}-\mathbf{EX})(\mathbf{X}-\mathbf{EX})^{\mathbf{T}}]\mathbf{B}^{\mathbf{T}} \\ &= \mathbf{B}\ \mathbf{Var}(\mathbf{X})\ \mathbf{B}^{\mathbf{T}} = \mathbf{B}\ \mathbf{\Sigma}\ \mathbf{B}^{\mathbf{T}} \end{aligned}$$

Actually, $\mathbf{y}$ must have a multivariate normal distribution with expected value $\mathbf{c}+\mathbf{B}\mu$ and variance $\mathbf{B\Sigma B^T}$ i.e., $\mathbf{y} \sim \mathcal{N}\left(\mathbf{c}+\mathbf{B}\boldsymbol{\mu}, \mathbf{B\Sigma B}^{\mathrm{T}}\right)$. The proof requires some advanced linear algebra and probability theory. If interested, please see the Appendix A.2 of multivariate gaussian worksheet on piazza.

2. Intercepts in Linear Regression

   In the traditional linear regression scenario, where we model $y$ with a line, or,

   $$\hat{y} = \vec{w}^T \vec{x}$$

   we aim to estimate $\vec{w}$. However, this model forces the lines to cross the origin (plug in $\vec{x} = \vec{0}$), severely limiting the power of the model. A typical solution to this problem is that the weight vector is extended by 1 and each input vector $\vec{x}$ has a 1 added at the beginning. This effectively adds an intercept term the model and allows for any line to be created.

   But that's boring! Let's come up with a solution to the intercept issue.

   Let's say we're stuck with the technique that fits only lines that go through the origin. We could shift the data to center it at zero, fit a line, and then shift our zero-centered line to where it's supposed to be. To center our data around the origin, we can subtract the mean of the $x$'s and the mean of the $y$'s from the data.

   a) Given that $\bar{x}$ and $\bar{y}$ are the means of our data, find the new model for a line predicting $y$ from $\vec{x}$.

   **Solution:** We shift our data by the means, centering it at 0. Then train a linear standard linear regression model on it. What is the new intercept term?

   $$y_{centered} = \vec{w}^T \vec{x}_{centered}$$

   $$y - \bar{y} = \vec{w}^T (\vec{x} - \bar{x})$$

   $$y = \vec{w}^T \vec{x} + (\bar{y} - \vec{w}^T \bar{x})$$

   The intercept is $\bar{y} - \vec{w}^T \bar{x}$.

   Another approach to the intercept term would be to just model an intercept in our equation and estimate it from the data. The model would now look like this:

   $$\hat{y} = \vec{w}^T \vec{x} + w_0$$

   b) Find the MLE estimate of $w_0$. You should get the same answer as before.

**Solution:** Assume that y is modeled with a line plus an intercept and Gaussian noise, or

$$y \sim \mathcal{N}(\vec{w}^T\vec{x} + w_0, \sigma^2)$$

We have $n$ training samples, $(x_1, x_2, x_3, \ldots, x_n)$. Performing MLE:

$$L(\vec{w}, w_0|x_1, \ldots, x_n) = P(x_1|\vec{w}, w_0)P(x_2|\vec{w}, w_0)\cdots P(x_n|\vec{w}, w_0)$$

$$L(\vec{w}, w_0|x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|\vec{w}, w_0)$$

$$L(\vec{w}, w_0|x_1, \ldots, x_n) = (\frac{1}{\sqrt{(2\pi\sigma^2)}})^n \exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\vec{w}^T\vec{x_i} + w_0))^2)$$

$$l(\vec{w}, w_0|x_1, \ldots, x_n) = n\ln(\frac{1}{\sqrt{(2\pi\sigma^2)}}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\vec{w}^T\vec{x_i} + w_0))^2$$

We want to maximize this the log likelihood, or equivalently minimize the following quantity:

$$\min(\frac{1}{2}\sum_{i=1}^{n}(y_i - (\vec{w}^T\vec{x_i} + w_0))^2)$$

$$\min(\frac{1}{2}\sum_{i=1}^{n}(y_i^2 - 2y_i(\vec{w}^T\vec{x_i} + w_0) + (\vec{w}^T\vec{x_i})^2 + 2w_0\vec{w}^T\vec{x_i} + w_0^2))$$

Taking the gradient with respect to $w_0$ gives us:

$$\frac{1}{2}\sum_{i=1}^{n}(-2y_i + 2w_0 + 2\vec{w}^T\vec{x_i}) = 0$$

$$\hat{w_0} = \frac{\sum_{i=1}^{n} y_i}{n} - \frac{\sum_{i=1}^{n} \hat{w}^T\vec{x_i}}{n}$$

3. Linearly Separable Data with Logistic Regression

   Show (or explain) that for a linearly separable data set, the maximum likelihood solution for the logistic regression model is obtained by finding a vector $\beta$ whose decision boundary $\beta^T x = 0$ separates the classes, and taking the magnitude of $\beta$ to be infinity.

**Solution:**

Because the data is linearly separable, it is possible to find a hyperplane with unit normal vector $\beta$ such that each halfspace induced by this hyperplane contain all samples of one class.

Consider all points on the half space defined by $\beta^T x \geq 0$. Without loss of generality, let's say that all these points come from class 1, while the points such that $\beta^T x < 0$ come from class -1. For some point $x_1$ in class 1,

$$P(y = 1|x_1) = \mu_i = \frac{1}{1 + exp(-\beta^T x_1)} > 0.5$$

because $\beta^T x_1 \geq 0$. Likewise, for a point $x_{-1}$ in class -1,

$$P(y = -1|x_{-1}) = 1 - P(y = 1|x_{-1}) = 1 - \mu_i > 0.5$$

since $\beta^T x_{-1} < 0$. Now, when we inspect the likelihood of the data, given by

$$L(\beta|D) = \prod_{i=1}^{n} \mu_i^{y_i}(1 - \mu_i)^{1-y_i} = \prod_{i \in w_1} \mu_i \prod_{j \in w_{-1}} (1 - \mu_j)$$

we see that if we take some arbitrary $c > 1$ and scale the unit vector $\beta$ by $c$, our likelihood will increase, since all of the individual probabilities in the likelihood will increase. In fact, we can set c = $\infty$, which will maximize our likelihood. This will render the sigmoid function to be infinitely steep at $\beta^T x_i = 0$ (making it a step function). $P(y = y_i|x_i) = 1$ for all $x_i$, and the likelihood will be 1. Obviously this is severely overfitting the data, and regularization for this problem would help us avoid that issue.