Problem Set 4 - TTIC

Ibrahim Gabr

Problem 1

$$\hat{p}(y=1|x;w,w_0) = \frac{1}{1+\exp\sum_{j=1}^{d} w_j\phi_j(\mathbf{x})}$$

Show how by appropriate choice of basis functions φ, given a training set $x_1, \ldots, x_N$, one can obtain a logistic regression model whose predictions on a test point x0 depend on the training data only through the kernel values $K(\mathbf{x_i}, \mathbf{x_0})$ for i = 1, . . . , N. Then write down the gradient of the loss, with L_2 regularization, on a single example, and show that the training for this model via gradient descent also depends on the training data only through kernel computations.

In order to apply the "kernel trick" to logistic regression, we can define our basis function as follows:

$$\phi_j(\mathbf{x}) = K(\mathbf{x}_j, \mathbf{x})$$

This basis function ϕ maps $\mathbf{x}$ from a low dimensional space X (dimensionality=d) into a high dimensional space Z (dimensonality is N, N > d), where each element after the constant represents the $\mathbf{x}_i$th Kernel product with the other sample's feature vectors.

$$\mathbf{x}_i \rightarrow \phi(\mathbf{x}_i) = [1, \phi_1(\mathbf{x}_i), \cdots, \phi_N(\mathbf{x}_i)]$$

Replacing this expression for the base function in the posterior probability we get:

$$\hat{p}(y=1|\mathbf{x};\mathbf{w},w_0) = \frac{1}{1+exp\left(\sum_{j=1}^{d} \mathbf{w}_j K(x_j,\mathbf{x})\right)}$$

Now, we can rewrite the log-loss function with L_2 reguralization:

$$log\, p(Y|X,\mathbf{w}) = \sum_{i=1}^{N} log\, p(y_i|\mathbf{x}_i,\mathbf{w}) + \lambda \sum_{j=1}^{d} w_j^2$$

Before computing the gradient, let's simplify the problem in the following way, and define $\xi = \sum_{j=1}^{d} w_j K(\mathbf{x}_j, \mathbf{x})$

$$\sum_{i=1}^{N} log\, p(y_i|\mathbf{x}_i, w_j) = \sum_{i=1}^{N} y_i\, log\left(\frac{1}{1+exp(\xi)}\right) + (1-y_i)\, log\left(1 - \frac{1}{1+exp(\xi)}\right)$$

$$= \sum_{i=1}^{N} -y_i\, log(1+exp(\xi)) + (1-y_i)\,(\xi - log(1+exp(\xi)))$$

$$= \sum_{i=1}^{N} \xi - y_i\xi - log(1+exp(\xi)) \quad \textbf{substituting} \quad \xi$$

$$= \sum_{i=1}^{N} \left(\sum_{j=1}^{d} w_j K(\mathbf{x}_j, \mathbf{x})\right) - y_i \left(\sum_{j=1}^{d} w_j K(\mathbf{x}_j, \mathbf{x})\right)$$

$$- log\left(1 + exp\left(\sum_{j=1}^{d} w_j K(\mathbf{x}_j, \mathbf{x})\right)\right)$$

For computing the gradient, it will be useful to compute in advance the derivate of ξ with respect to w_j:

$$\frac{\delta \xi}{\delta w_j} = K(\mathbf{x}_j, \mathbf{x})$$

Now, we are in the position to state the minimization problem as follows:

$$log\, p(Y|X, \mathbf{w}) = \sum_{i=1}^{N} \left(\sum_{j=1}^{d} w_j K(\mathbf{x}_j, \mathbf{x})\right) - y_i \left(\sum_{j=1}^{d} w_j K(\mathbf{x}_j, \mathbf{x})\right)$$

$$- log\left(1 + exp\left(\sum_{j=1}^{d} w_j K(\mathbf{x}_j, \mathbf{x})\right)\right) + \lambda \sum_{j=1}^{d} w_j^2$$

$$\frac{\delta L}{\delta w_j} = \sum_{i=1}^{N} K(\mathbf{x}_j, \mathbf{x}) - y_i K(\mathbf{x}_j, \mathbf{x}) - \frac{K(\mathbf{x}_j, \mathbf{x})\, exp\left(\sum_{j=1}^{d} w_j K(\mathbf{x}_j, \mathbf{x})\right)}{1 + exp\left(\sum_{j=1}^{d} w_j K(\mathbf{x}_j, \mathbf{x})\right)} + 2\lambda w_j$$

$$= \sum_{i=1}^{N} K(\mathbf{x}_j, \mathbf{x})(1-y_i) - \frac{K(\mathbf{x}_j, \mathbf{x})\, exp\left(\sum_{j=1}^{d} w_j K(\mathbf{x}_j, \mathbf{x})\right)}{1 + exp\left(\sum_{j=1}^{d} w_j K(\mathbf{x}_j, \mathbf{x})\right)} + 2\lambda w_j$$

Therefore, the equation for the stochastic gradient descent depends on the training data only through Kernel computations.

$$w_j^{(t+1)} = w_j^{(t)} + \eta \frac{\delta L}{\delta w_j}$$

Problem 2

The discriminant function for Gaussians with equal covariances, $\Sigma_c = \Sigma \quad \forall c$, can be characterized as follows:

$$\begin{aligned}
\delta_c(x) &= log\, p(\mathbf{x}|y=c) \\
&= -log(2\pi)^{d/2} - \frac{1}{2} log(|\Sigma|) - \frac{1}{2}(\mathbf{x}-\mu_c)^T \Sigma^{-1}(\mathbf{x}-\mu_c) \\
&= -log(2\pi)^{d/2} - \frac{1}{2} log(|\Sigma|) - \frac{1}{2}\left(\mathbf{x}^T\Sigma^{-1}\mathbf{x} - \mathbf{x}^T\Sigma^{-1}\mu_c - \mu_c\Sigma^{-1}\mathbf{x} + \mu_c^T\Sigma^{-1}\mu_c\right) \\
&= \mu_c^T\Sigma^{-1}\mathbf{x} - \frac{1}{2}\mu_c^T\Sigma_c^{-1}\mu_c - \text{const}_c(\mathbf{x}) \\
\text{where} \quad \text{const}_c(\mathbf{x}) &= -log(2\pi)^{d/2} - \frac{1}{2} log(|\Sigma|) - \frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}
\end{aligned}$$

From Bayes rule, we know that $p(\mathbf{x}|y=c)$ can be expressed as:

$$p(\mathbf{x}|y=c) = \frac{p(y=c|\mathbf{x})p(\mathbf{x})}{p(y=c)}$$

Therefore, the posterior probability $p(y=c|\mathbf{x})$ is given by:

$$
\begin{aligned}
p(y=c|\mathbf{x}) &= \frac{p(y=c)p(\mathbf{x}|y=c)}{p(\mathbf{x})} \\
&= \frac{p(y=c)p(\mathbf{x}|y=c)}{\sum_c p(\mathbf{x}, y=c)} \\
&= \frac{p(y=c)p(\mathbf{x}|y=c)}{\sum_c p(y=c)p(\mathbf{x}|y=c)} \\
&= \frac{p(\mathbf{x}|y=c)}{\sum_c p(\mathbf{x}|y=c)}
\end{aligned}
$$

$$
\begin{aligned}
\textbf{where} \quad p(\mathbf{x}, y=c) &= p(y=c|\mathbf{x})p(\mathbf{x}) \\
&= \frac{p(y=c)p(\mathbf{x}|y=c)}{p(\mathbf{x})}p(\mathbf{x}) \\
&= p(y=c)p(\mathbf{x}|y=c)
\end{aligned}
$$

Now, we need to take the exponential from $log\, p(\mathbf{x}|y=c)$ to get $p(\mathbf{x}|y=c)$.

$$p(\mathbf{x}|y=c) = exp(\delta_c(\mathbf{x}))$$

Also, we can express the difference between two class discriminants as:

$$
\begin{aligned}
\delta_{c_2}(\mathbf{x}) - \delta_{c_1}(\mathbf{x}) &= -\frac{1}{2}\mu_{c_2}^T \Sigma^{-1} \mu_{c_2} + \mu_{c_2}^T \Sigma^{-1}\mathbf{x} - \text{const}_{c_2}(\mathbf{x}) \\
&\quad - -\frac{1}{2}\mu_{c_1}^T \Sigma^{-1} \mu_{c_1} + \mu_{c_1}^T \Sigma^{-1}\mathbf{x} - \text{const}_{c_1}(\mathbf{x}) \\
&= \mathbf{w} \cdot \mathbf{x} + w_0
\end{aligned}
$$

Given that we only have to classes, we can express the posterior probability of c_1 as follows:

$$
\begin{aligned}
p(y=c_1|\mathbf{x}) &= \frac{p(\mathbf{x}|y=c_1)}{\sum_{k=1}^2 p(\mathbf{x}|y=k)} \\
&= \frac{exp(\delta_{c_1}(\mathbf{x}))}{\sum_{k=1}^2 exp(\delta_{c_k}(\mathbf{x}))} \\
&= \frac{1}{1 + exp(\delta_{c_2}(\mathbf{x}) - \delta_{c_1}(\mathbf{x}))} \\
&= \frac{1}{1 + exp(\mathbf{w} \cdot \mathbf{x} + w_0)}
\end{aligned}
$$

Problem 3

he logistic regression model and the linear discriminant analysis (LDA) based on the isotropic Gaussian model should give similar classifiers.

The difference between linear logistic regression and LDA is that the linear logistic model only specifies the conditional distribution $P(Y = c|X = x)$. No assumption is made about $P(X)$; while the LDA model specifies the joint distribution of X and Y. In LDA, $P(X)$ is a mixture of Gaussians.

Therefore, LDA gives better results if the underlying data X follows a multivariate Gaussian distribution. Since the Logistic regression makes no assumptions about the distribution of the data, it is more general than LDA, and will give similar results even if the data follow a multivariate Gaussian distribution.

Furthermore, even if the true $f_k(x)$ are gaussian distributed, employing logistic regression will result in an an efficiency loss of 30% asymptotically in the error. In other words, gathering 30% more data will result will result in a conditional likelihood that will perform just as well as the LDA. (Trevor J. Hastie, Robert John Tibshirani, and Jerome H. Friedman, *The elements of statistical learning: data mining, inference, and prediction* 2001.)

To conclude, the classifiers will be similar.