Problem Set 2

Problem 2.1 (Bayesian analysis of the exponential distribution)

A lifetime X of a machine is modeled by an exponential distribution with unknown parameter θ. The likelihood is $p(x \mid \theta) = \theta e^{-\theta x}$ for $x \geq 0, \theta > 0$.

(a) Show that the MLE is $\hat{\theta} = 1/\bar{x}$, where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$.

(b) Suppose we observe $X_1 = 5, X_2 = 6, X_3 = 4$ (the lifetimes (in years) of 3 different iid machines). What is the MLE given this data?

(c) Assume that an expert believes θ should have a prior

$$p(\theta) = \text{Expon}(\theta \mid \lambda)$$

Choose the prior parameter, call it $\hat{\lambda}$, such that $\mathbb{E}[\theta] = 1/3$. Hint: recall that the Gamma distribution has the form

$$\text{Ga}(\theta \mid a, b) \propto \theta^{a-1} e^{-\theta b}$$

and its mean is a/b.

(d) What is the posterior, $p(\theta \mid \mathcal{D}, \hat{\lambda})$?

(e) Is the exponential prior conjugate to the exponential likelihood?

(f) What is the posterior mean, $\mathbb{E}[\theta \mid \mathcal{D}, \hat{\lambda}]$?

(g) Explain why the MLE and posterior mean differ. Which is more reasonable in this example?

Problem 2.2 (Sufficient statistics for online linear regression)

Consider fitting the model $\hat{y} = w_0 + w_1 x$ using least squares. Unfortunately we did not keep the original data, x_i, y_i, but we do have the following functions (statistics) of the data:

$$\begin{aligned} \bar{x}^{(n)} &= \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y}^{(n)} = \frac{1}{n}\sum_{i=1}^{n} y_i \\ C_{xx}^{(n)} &= \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2, \quad C_{xy}^{(n)} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}), \quad C_{yy}^{(n)} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})^2 \end{aligned} \tag{1}$$

(a) What are the minimal set of statistics that we need to estimate w_1?

(b) What are the minimal set of statistics that we need to estimate w_0?

(c) Suppose a new data point, x_{n+1}, y_{n+1} arrives, and we want to update our sufficient statistics without looking at the old data, which we have not stored. (This is useful for online learning.) Show that we can do this for $\bar{x}$ as follows.

$$\begin{aligned}\bar{x}^{(n+1)} &\triangleq \frac{1}{n+1}\sum_{i=1}^{n+1} x_i = \frac{1}{n+1}\left(n\bar{x}^{(n)} + x_{n+1}\right) \\ &= \bar{x}^{(n)} + \frac{1}{n+1}\left(x_{n+1} - \bar{x}^{(n)}\right)\end{aligned}$$

This has the form: new estimate is old estimate plus correction. We see that the size of the correction diminishes over time (i.e., as we get more samples). Derive a similar expression to update $\bar{y}$.

(d) Show that one can update $C_{xy}^{(n+1)}$ recursively using

$$C_{xy}^{(n+1)} = \frac{1}{n+1}\left[x_{n+1}y_{n+1} + nC_{xy}^{(n)} + n\bar{x}^{(n)}\bar{y}^{(n)} - (n+1)\bar{x}^{(n+1)}\bar{y}^{(n+1)}\right]$$

Derive a similar expression to update C_{xx}.

Problem 2.3 (Symmetric version of ℓ_2 regularized multinomial logistic regression)

Multiclass logistic regression has the form

$$p(y = c \mid \mathbf{x}, \mathbf{W}) = \frac{\exp\left(w_{c0} + \mathbf{w}_c^T\mathbf{x}\right)}{\sum_{k=1}^{C}\exp\left(w_{k0} + \mathbf{w}_k^T\mathbf{x}\right)} \tag{2}$$

where $\mathbf{W}$ is a $(D+1)\times C$ weight matrix. We can arbitrarily define $\mathbf{w}_c = \mathbf{0}$ for one of the classes, say $c = C$, since $p(y = C \mid \mathbf{x}, \mathbf{W}) = 1 - \sum_{c=1}^{C-1} p(y = c \mid \mathbf{x}, \mathbf{w})$. In this case, the model has the form

$$p(y = c \mid \mathbf{x}, \mathbf{W}) = \frac{\exp\left(w_{c0} + \mathbf{w}_c^T\mathbf{x}\right)}{1 + \sum_{k=1}^{C-1}\exp\left(w_{k0} + \mathbf{w}_k^T\mathbf{x}\right)}$$

If we don't "clamp" one of the vectors to some constant value, the parameters will be unidentifiable. However, suppose we don't clamp $\mathbf{w}_c = \mathbf{0}$, so we are using 2, but we add ℓ_2 regularization by optimizing

$$\sum_{i=1}^{N}\log p\left(y_i \mid \mathbf{x}_i, \mathbf{W}\right) - \lambda\sum_{c=1}^{C}\|\mathbf{w}_c\|_2^2$$

Show that at the optimum we have $\sum_{c=1}^{C}\hat{w}_{cj} = 0$ for $j = 1 : D$. (For the unregularized $\hat{w}_{c0}$ terms, we still need to enforce that $w_{0C} = 0$ to ensure identifiability of the offset.)

Problem 2.4 (SVM with Gaussian kernel)

Consider the task of training a support vector machine using the Gaussian kernel $K(x, z) = \exp\left(-\|x - z\|^2/\tau^2\right)$. We will show that as long as there are no two identical points in the training set, we can always find a value for the bandwidth parameter τ such that the SVM achieves zero training error.

(a) Recall from class that the decision function learned by the support vector machine can be written as

$$f(x) = \sum_{i=1}^{m} \alpha_i y^{(i)} K\left(x^{(i)}, x\right) + b.$$

Assume that the training data $\left\{\left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(m)}, y^{(m)}\right)\right\}$ consists of points which are separated by at least a distance of ϵ; that is, $\left\|x^{(j)} - x^{(i)}\right\| \geq \epsilon$ for any $i \neq j$. Find values for the set of parameters $\{\alpha_1, \ldots, \alpha_m, b\}$ and Gaussian kernel width τ such that $x^{(i)}$ is correctly classified, for all $i = 1, \ldots, m$. [Hint: Let $\alpha_i = 1$ for all i and $b = 0$. Now notice that for $y \in \{-1, +1\}$ the prediction on $x^{(i)}$ will be correct if $\left|f\left(x^{(i)}\right) - y^{(i)}\right| < 1$, so find a value of τ that satisfies this inequality for all i.]

(b) Suppose we run a SVM with slack variables using the parameter τ you found in part (a). Will the resulting classifier necessarily obtain zero training error? Why or why not? A short explanation (without proof) will suffice.