This handout includes space for every question that requires a written response. Please feel free to use it to handwrite your solutions (legibly, please). If you choose to typeset your solutions, the `README.md` for this assignment includes instructions to regenerate this handout with your typeset LaTeX solutions.

1.a

The distribution is in the exponential family if we could write it as:

$$p(y;\eta) = b(y)\exp(\eta T(y) - a(\eta)) \tag{1}$$

Let's rewrite the Poisson distribution to this form:

$$p(y;\lambda) = \frac{1}{y!}\lambda^y \exp(-\lambda) = \frac{1}{y!}\exp(\log\lambda^y)\exp(-\lambda) = \frac{1}{y!}\exp(-\lambda + y\log\lambda)$$

Let's set $\eta = \log\lambda$, then $\lambda = \exp(\eta)$:

$$p(y;\eta) = \frac{1}{y!}\exp(y\eta - \exp(\eta))$$

So let's match and write the parameters of exponential family as in formula (1):

$$\begin{aligned} b(y) &= \frac{1}{y!} \\ \eta &= \log\lambda \\ T(y) &= y \\ a(\eta) &= \exp(\eta) \end{aligned}$$

1.b

For the Poisson distribution mean of the response $\mathbb{E}[Y] = \lambda$. Canonical link function as shown in 1.a. is:

$$\eta = \log \lambda$$

Therefore, the canonical response function is:

$$g^{-1}(\eta) = \lambda = \exp(\eta)$$

1.c

The log-likelihood of an example $(x^{(i)}, y^{(i)})$ is defined as $\ell(\theta) = \log p(y^{(i)}|x^{(i)};\theta)$. To derive the stochastic gradient ascent rule, use the results in part (a) and the standard GLM assumption that $\eta = \theta^T x$.

$$\begin{aligned}
\frac{\partial \ell(\theta)}{\partial \theta_j} &= \frac{\partial \log p(y^{(i)}|x^{(i)};\theta)}{\partial \theta_j} = \\
&= \frac{\partial \log\left(\frac{1}{y^{(i)}!}\exp(\eta^T y^{(i)} - e^{\eta})\right)}{\partial \theta_j} = \\
&= \cancel{\frac{\partial \log \frac{1}{y^{(i)}!}}{\partial \theta_j}} + \frac{\partial(\eta^T y^{(i)} - e^{\eta})}{\partial \theta_j} = \\
&= \frac{\partial(\theta^T x^{(i)} y^{(i)} - e^{\theta^T x^{(i)}})}{\partial \theta_j} = \\
&= x_j^{(i)}(y^{(i)} - exp(\theta^T x^{(i)}))
\end{aligned}$$

Thus the stochastic gradient ascent update rule should be:

$$\theta_j := \theta_j + \alpha \frac{\partial \ell(\theta)}{\partial \theta_j}$$

which reduces here to:

$$\theta_j := \theta_j + \alpha x_j^{(i)}(y^{(i)} - exp(\theta^T x^{(i)}))$$

2.a

For a new kernel K:

$$K(x,z) = K_1(x,z) + K_2(x,z)$$

We need to find out if $K(x,z)$ is a kernel - i.e. is symmetric and positive semidefinite square matrix. We know that $K_1 \in \mathbb{R}^{n\times n}$, $K_2 \in \mathbb{R}^{n\times n}$ are symmetric and positive semidefinite square matrices. Therefore:

$$x^T K_1 x \geq 0, \forall x \in \mathbb{R}^d$$
$$x^T K_2 x \geq 0, \forall x \in \mathbb{R}^d$$

Consequently, for the matrix $K = K_1 + K_2$ the following is valid:

$$x^T K x = x^T (K_1 + K_2) x = x^T K_1 x + x^T K_2 x$$

As both $x^T K_1 x \geq 0$ and $x^T K_2 x \geq 0$, then $\forall x \in \mathbb{R}^n$:

$$x^T K x = x^T K_1 x + x^T K_2 x \geq 0$$

Since the sum of symmetric square matrices is also symmetric square matrix of the same dimensions we make a conclusion that $K = K_1 + K2$, $K \in \mathbb{R}^{n\times n}$ is a valid kernel.

2.b

For a new kernel K:

$$K(x,z) = K_1(x,z) - K_2(x,z)$$

We need to find out if $K(x,z)$ is a kernel - i.e. is symmetric and positive semidefinite square matrix. We know that $K_1 \in \mathbb{R}^{n\times n}$, $K_2 \in \mathbb{R}^{n\times n}$ are symmetric and positive semidefinite square matrices. Therefore:

$$x^T K_1 x \geq 0, \forall x \in \mathbb{R}^d$$
$$x^T K_2 x \geq 0, \forall x \in \mathbb{R}^d$$

Let's choose $K2{=}{-}2K_1$. Under the new circumstances we have the following equality:

$$K(x,z) = K_1(x,z) - 2K_1(x,z) = -K_1(x,z)$$

Since K_1 is positive semidefinite and has positive eigenvalues then $-K_1$ is negative semidefinite since it has negative eigenvalues. So the difference of 2 positive semidefinite kernels is not necessarily a valid kernel.

2.c

For a new kernel K:

$$K(x, z) = \alpha K_1(x, z)$$

We need to find out if $K(x, z)$ is a kernel - i.e. is symmetric and positive semidefinite square matrix. We know that $K_1 \in \mathbb{R}^{n \times n}$ is symmetric and positive semidefinite square matrix and $\alpha \in \mathbb{R}^+$. Therefore:

$$x^T K x = x^T(\alpha K_1)x = \alpha x^T K_1 x$$

Since $\alpha \in \mathbb{R}^+$ and K_1 positive semidefinite, then $\alpha x^T K_1 x \geq 0$ and K is positive semidefinite, is a valid kernel.

2.d

For a new kernel K:

$$K(x,z) = -\alpha K_1(x,z)$$

We need to find out if $K(x,z)$ is a kernel - i.e. is symmetric and positive semidefinite square matrix. We know that $K_1 \in \mathbb{R}^{n\times n}$ is symmetric and positive semidefinite square matrix and $\alpha \in \mathbb{R}^+$. Therefore:

$$x^T K x = x^T(-\alpha K_1)x = -\alpha x^T K_1 x$$

Since $\alpha \in \mathbb{R}^+$ and K_1 positive semidefinite, then $-\alpha x^T K_1 x \leq 0$ and K is negative semidefinite, i.e. is not a valid kernel.

3.ai

So, we have a kernel K that corresponds to feature mapping ϕ to start with. Also we need to change the rule for update of $\theta^{(i+1)}$ from original dimensionality to high-dimensional feature space ϕ:

$$\theta^{(i+1)} := \theta^{(i)} + \alpha(y^{(i+1)} - h_{\theta^{(i)}}(x^{(i+1)}))x^{(i+1)}$$

to:

$$\theta^{(i+1)} := \theta^{(i)} + \alpha y^{(i+1)}\phi(x^{(i+1)}) = \theta^{(i)} + \beta_{(i+1)}\phi(x^{(i+1)})$$

In perceptron for each new $i+1$ we make $\beta_{i+1} = 0$ if prediction is correct and corresponds to $y^{(i+1)}$. If it doesn't correspond, we assign $\beta_{i+1} = \alpha y^{(i+1)}$ if $y^{(i+1)} \neq h_{\theta^{(i)}}(x^{(i+1)})$.

We start with initialization of $\theta = 0$ and focus on iterative update. At any time θ is represented as a linear combination of vectors $\phi(x^{(i)}), ..., \phi(x^{(n)})$.

Therefore $\theta^{(i)}$ is being updated by multiple of $\phi(x^{(i+1)})$. An expression for θ at some point is a linear combination of past $\phi(x^{(j)})$:

$$\theta^{(i)} = \sum_{j=1}^{i} \beta_j \phi(x^{(j)})$$

Let's write an expression for h for a new $x^{(i+1)}$:

$$\theta^{(i)^T}\phi(x^{(i+1)}) = \sum_{j=1}^{i} \beta_j \phi(x^{(j)})^T \phi(x^{(i+1)}) = \sum_{j=1}^{i} \beta_j K(x^{(j)}, x^{(i+1)})$$

3.aii

Let's find the expression for $h_{\theta^{(i)}}(x^{(i+1)})$. The result of the previous task - for each new x:

$$\theta^T \phi(x) = \sum_{i=1}^{n} \beta_i \phi(x^{(i)})^T \phi(x) = \sum_{i=1}^{n} \beta_i K(x^{(i)}, x)$$

Therefore, for $x^{(i+1)}$:

$$\theta^T \phi(x^{(i+1)}) = \sum_{j=1}^{i} \beta_i \phi(x^{(j)})^T \phi(x^{(i+1)}) = \sum_{j=1}^{i} \beta_i K(x^{(j)}, x^{(i+1)})$$

Consequently:

$$h_{\theta^{(i)}}(x^{(i+1)}) = g(\theta^T \phi(x^{(i+1)})) = g(\sum_{j=1}^{i} \beta_i K(x^{(j)}, x^{(i+1)}))$$

And finally:

$$h_{\theta^{(i)}}(x^{(i+1)}) = \begin{cases} 1, & \text{if } g(\sum_{j=1}^{i} \beta_i K(x^{(j)}, x^{(i+1)})) \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

3.aiii

The prediction on new input $(x^{(i+1)}, y^{(i+1)})$ will be:

$$\theta^{(i+1)} = \theta^{(i)} + \alpha(y^{(i+1)} - h_{\theta^{(i)}}(x^{(i+1)}))\phi(x^{(i+1)}) = \theta^{(i)} + \beta_{i+1}\phi(x^{(i+1)})$$

3.c

The perceptron classifier plot for dot-product kernel performs worse than one with radial basis function kernel on the given data, because the first one learns the linear decision boundary. While the data itself lies in radial clusters. For this type of data the radial basis function kernel is a better choice since it could capture such kind of nonlinearity and better suited for radially structured data.