## Problem 1 (10 marks)

**(A Circular Definition)** Consider a logistic regression model $p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \frac{1}{1+\exp(-y_n \boldsymbol{w}^\top \boldsymbol{x}_n)}$, with a zero-mean Gaussian prior $p(\boldsymbol{w}) = \mathcal{N}(\mathbf{0}, \lambda^{-1}\mathbf{I})$. Note that this loss function for logistic regression assumes $y_n \in \{-1, +1\}$ instead of $\{0, 1\}$. Show that the MAP estimate for $\boldsymbol{w}$ can be written as $\boldsymbol{w} = \sum_{n=1}^{N} \alpha_n y_n \boldsymbol{x}_n$ where each $\alpha_n$ itself is a function of $\boldsymbol{w}$. Based on the expression of $\alpha_n$, you would see that it has a precise meaning. Briefly state (in 50 words of less, may include equations) what $\alpha_n$ means, and also briefly explain (in words or less) why the result $\boldsymbol{w} = \sum_{n=1}^{N} \alpha_n y_n \boldsymbol{x}_n$ makes sense for this model.

## Problem 2 (10 marks)

**(Generative meets Discriminative)** Consider a generative classification model for binary classification. Assume the class-marginal distribution to be defined as $p(y = 1) = \pi$ and assume each class-conditional distribution to be defined as a product of $D$ Bernoulli distributions, i.e., $p(\boldsymbol{x}|y = 1) = \prod_{d=1}^{D} p(x_d|y = 1)$ where $p(x_d|y = 1) = \text{Bernoulli}(x_d|\mu_{d,1})$, and $p(\boldsymbol{x}|y = 0) = \prod_{d=1}^{D} p(x_d|y = 0)$ where $p(x_d|y = 0) = \text{Bernoulli}(x_d|\mu_{d,0})$. Note that this makes uses of the naïve Bayes assumption.

Show that this model is equivalent (in its mathematical form) to a probabilistic discriminative classifier. In particular, derive the expression for $p(y = 1|\boldsymbol{x})$, and state what type of decision boundary will this model learn - linear, quadratic, or something else (looking at the expression of $p(y = 1|\boldsymbol{x})$ should reveal that)? Clearly write down the expressions for the parameters of the equivalent probabilistic discrimininative model in terms of the generative model parameters $(\pi, \mu_{d,0}, \mu_{d,1})$. Note that you do not have to estimate the parameters $\pi, \mu_{d,0}, \mu_{d,1}$ (but you may try that for practice if you want).

## Problem 3 (5 marks)

**(The Equivalence)** Consider a constrained version of least squares linear regression where we constrain the $\ell_2$ norm of $\boldsymbol{w}$ to be less than on equal to some $c > 0$: $\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2, \quad \text{s.t.} \quad ||\boldsymbol{w}|| \leq c$

Show, formally, that it is possible to have an $\ell_2$ regularized least squares linear regression model that will give the *exact* same solution as the solution to the above constrained problem.

## Problem 4 (20 marks)

**(Softmax and Variants)** Consider $N$ training examples $\{\boldsymbol{x}_n, y_n\}_{n=1}^{N}$ where $\boldsymbol{x}_n \in \mathbb{R}^D$, and $y_n \in \{1, \ldots, K\}$. Suppose we wish to use this data to learn the multiclass logistic (or "softmax") regression model which defines $p(y_n = k|\boldsymbol{x}_n, \mathbf{W}) = \mu_{nk} = \frac{\exp(\boldsymbol{w}_k^\top \boldsymbol{x}_n)}{\sum_{\ell=1}^{K} \exp(\boldsymbol{w}_\ell^\top \boldsymbol{x}_n)}$, where $\mu_{nk}$ is the predicted probability of $y_n = k$.

Derive the MLE solution for $\mathbf{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_K]$. You would notice that, just like logistic regression, there is no closed form solution for $\mathbf{W}$. So you will need to write down the log likelihood, take its derivative w.r.t. each column $\boldsymbol{w}_k$ of $\mathbf{W}$ to compute the gradient, and derive the gradient descent (GD) update rule for each $\boldsymbol{w}_k$. Show the basic steps of your derivation and write down the final expression for the GD update of each $\boldsymbol{w}_k$. Assume a fixed learning rate $\eta = 1$ for simplicity.

Next, consider the stochastic gradient descent (SGD) update for the same model where each iteration takes a randomly chosen example $(\boldsymbol{x}_n, y_n)$ to update $\mathbf{W}$. Write down the expressions for these SGD updates and the overall sketch of the corresponding SGD algorithm.

Finally consider a special case of the above SGD algorithm for this model, where the predicted "soft" class probabilities $\mu_{nk}$ are replaced by hard class assignments, i.e., $\mu_{nk} = 1$ for $k = \arg\max_\ell \{\mu_{n\ell}\}_{\ell=1}^{K}$, and $\mu_{nk'} = 0, \forall k' \neq k$. Write down the expressions for the SGD update in this case and the overall sketch of the SGD algorithm. How are these updates different from the previous case where you used soft probabilities $\mu_{nk}$?

## Problem 5 (10 marks)

**(Separating Convex Hulls)** Given a set of data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, we define the convex hull to be the set of all points $\boldsymbol{x}$ given by $\boldsymbol{x} = \sum_n \alpha_n \boldsymbol{x}_n$ where $\alpha_n \geq 0$ and $\sum_n \alpha_n = 1$ (Intuitively, the convex hull of a set of points in the solid region that they enclose.) Consider a second set of points $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M$ together with their corresponding convex hull. Show that the set of $\boldsymbol{x}$s and the set of $\boldsymbol{y}$s are linearly separable *if and only if* the convex hulls do not intersect.

## Problem 6 (5 marks)

**(Arbitrary Choice?)** Formally, show that changing the condition $y_n(\boldsymbol{w}^\top \boldsymbol{x}_n + b) \geq 1$ in SVM to a different condition $y_n(\boldsymbol{w}^\top \boldsymbol{x}_n + b) \geq m$ does not change the effective separating hyperplane that is learned by the SVM. Assume the hard-margin SVM for simplicity.