Name:

Roll No.: **Dept.:**

IIT Kanpur
CS771 Intro to ML
Mid-semester Examination
Date: September 21, 2017

Instructions:

Total: **80 marks**

Problem 1 (True or False: 8 X 1 = 8 marks). For each of the following simply write **T** or **F** in the box.

1. **F** The Bayesian predictive posterior has a nice closed form solution if we have a logistic likelihood for $\mathbb{P}[y \mid \mathbf{x}, \mathbf{w}]$ and a Gaussian prior for $\mathbb{P}[\mathbf{w}]$. (In fact even the MAP estimate does not have a closed solution and Laplace and other approximations need to be used.)
2. **F** Hard assignment alternating optimization approaches are much more expensive to execute than soft assignment alternating optimization approaches. (Hard assignment is usually cheaper)
3. **F** In ridge regression ($\arg\min \lambda/2 \cdot \|\mathbf{w}\|_2^2 + \left\|X^\top \mathbf{w} - \mathbf{y}\right\|_2^2$), no matter how large a regularization constant $\lambda > 0$ we set, we will always get good solutions. (An extremely large regularization will yield a trivial solution)
4. **T** When deriving MLE solutions, working with log-likelihood terms is simpler than working with likelihood terms directly. (Yes, it reduces product terms to sum terms that are easier to optimize)
5. **F** It is okay to perform minor evaluations on the test set during training so long as we don't do it too many times. (It is illegal to look at the test set for *anything* during training)
6. **F** If S_1 and S_2 are two convex sets in $\mathbb{R}^2$, then their union $S_1 \cup S_2$ is always a convex set as well. (Take 2 disjoint rectangles. They are individually convex but their union is not even a connected set.)
7. **F** It is not possible to execute the SGD algorithm if the objective function is not differentiable. (It is possible to use subgradients to execute the SGD algorithm even with non-differentiable losses like the hinge loss)
8. **T** Convex optimization problems like ridge regression are not as sensitive to proper initialization (while carrying out optimization) as are non-convex problems like k-means. (There is some sensitivity but the dependence is not nearly as chaotic as in case of k-means)

Problem 2 (Ultra Short Answer: 6 x 4 = 24 marks). Give your answers in the space provided only.

1. Suppose I have a coin with bias p i.e. it lands heads with probability p. What is the probability that when this coin is tossed n times, we observe x heads and $n - x$ tails? Give only the final expression.

   Solution: $\frac{n!}{x!(n-x)!} \cdot p^x (1-p)^{n-x}$

2. Given a vector $\mathbf{a} \in \mathbb{R}^d$, what is the trace of the matrix $A = \mathbf{a}\mathbf{a}^\top \in \mathbb{R}^{d \times d}$?

   Solution: $\operatorname{tr}(A) = \|\mathbf{a}\|_2^2$

3. Give the time complexity of predicting the label of a new point using the OvA and AvA approaches in a multiclassification problem with K classes with d-dimensional features. Briefly justify your answer.

   Solution: Time complexity for OvA is $\mathcal{O}(d \cdot k)$ since k models have to be evaluated and each evaluation takes $\mathcal{O}(d)$ time. Time complexity for AvA is $\mathcal{O}(d \cdot k^2)$ since $k(k-1)$ models have to be evaluated.

4. We are given that $\mathbb{P}[\boldsymbol{\Theta}] = 0.1, \mathbb{P}[y \mid \mathbf{x}, \boldsymbol{\Theta}] = 0.4, \mathbb{P}[\mathbf{x} \mid y, \boldsymbol{\Theta}] = 0.5, \mathbb{P}[y \mid \boldsymbol{\Theta}] = 0.2, \mathbb{P}[\mathbf{x}, y] = 0.5$.
   Find $\mathbb{P}[\boldsymbol{\Theta} \mid \mathbf{x}, y]$ and $\mathbb{P}[\mathbf{x} \mid \boldsymbol{\Theta}]$. Show your expressions for these terms briefly and the final answer

Solution: We have

$$\mathbb{P}\left[\boldsymbol{\Theta} \,|\, \mathbf{x}, y\right] = \frac{\mathbb{P}\left[\boldsymbol{\Theta}, \mathbf{x}, y\right]}{\mathbb{P}\left[\mathbf{x}, y\right]} = \frac{\mathbb{P}\left[\mathbf{x} \,|\, y, \boldsymbol{\Theta}\right] \cdot \mathbb{P}\left[y \,|\, \boldsymbol{\Theta}\right] \cdot \mathbb{P}\left[\boldsymbol{\Theta}\right]}{\mathbb{P}\left[\mathbf{x}, y\right]} = \frac{0.5 \cdot 0.2 \cdot 0.1}{0.5} = 0.02$$

We similarly have

$$0.4 = \mathbb{P}\left[y \,|\, \mathbf{x}, \boldsymbol{\Theta}\right] = \frac{\mathbb{P}\left[y, \mathbf{x} \,|\, \boldsymbol{\Theta}\right]}{\mathbb{P}\left[\mathbf{x} \,|\, \boldsymbol{\Theta}\right]} = \frac{\mathbb{P}\left[\mathbf{x} \,|\, y, \boldsymbol{\Theta}\right] \cdot \mathbb{P}\left[y \,|\, \boldsymbol{\Theta}\right]}{\mathbb{P}\left[\mathbf{x} \,|\, \boldsymbol{\Theta}\right]} = \frac{0.5 \cdot 0.2}{\mathbb{P}\left[\mathbf{x} \,|\, \boldsymbol{\Theta}\right]},$$

which gives us

$$\mathbb{P}\left[\mathbf{x} \,|\, \boldsymbol{\Theta}\right] = \frac{0.5 \cdot 0.2}{0.4} = 0.25$$

5. Consider a regression problem with covariates $\mathbf{x}^i \in \mathbb{R}^d$ and responses $y^i \sim \mathcal{N}(\langle \mathbf{w}, \mathbf{x}^i \rangle, \sigma^2)$. Suppose you are given $(\mathbf{x}^i, y^i)_{i=1,2,\ldots,n}$ as well as $\mathbf{w}$. Write down an estimator for σ.

Solution: The maximum likelihood estimate is (derivation similar to that in [**DAU**] S. 9.5)

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2}$$

6. Let $\mathbf{z}^t \in [K]^n$ denote the cluster assignments made by the k-means algorithm at the t-th iteration i.e. data point $i \in [n]$ gets assigned to the cluster $\mathbf{z}^t_i \in [K]$. Suppose we have $\mathbf{z}^t \neq \mathbf{z}^{t+1}$ but $\mathbf{z}^t = \mathbf{z}^{t'}$ for some $t' > t + 1$? What must be happening if cluster assignments get repeated in this manner?

Solution: Note that the k-means algorithm does not let the k-means objective (lec12.pdf page 147) value go up at any alternation

(a) It fixes the cluster assignments and finds the best cluster centers for this assignment which cannot let the objective value go up.

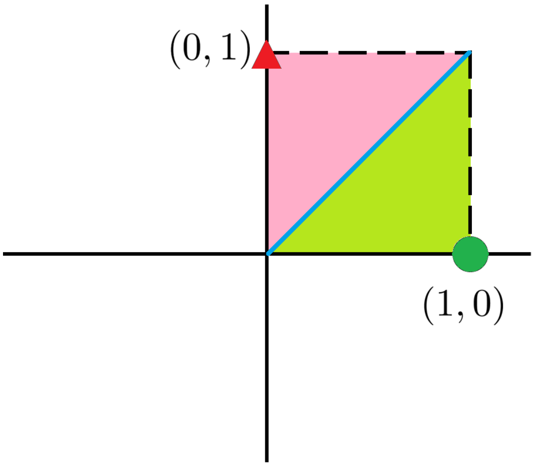
(b) It fixes the centers and finds the best cluster assignments for these centers which also cannot let the objective value go up.

Thus, if cluster assignments repeat non-trivially then it means that if we take the cluster assignments $\mathbf{z}^t, \mathbf{z}^{t+1}, \ldots, \mathbf{z}^{t'} = \mathbf{z}^t$, and for each of them compute cluster centers $\left\{\boldsymbol{\mu}^{t,k}\right\}, \left\{\boldsymbol{\mu}^{t+1,k}\right\}, \ldots, \left\{\boldsymbol{\mu}^{t',k}\right\} = \left\{\boldsymbol{\mu}^{t,k}\right\}$ using step 3 of the k-means algorithm, then all the pairings $\left\{\mathbf{z}^\tau, \left\{\boldsymbol{\mu}^{\tau,k}\right\}\right\}$ for $\tau = t, t+1, \ldots, t'$ offer the same k-means objective value. Essentially the algorithm has gotten trapped at a local minimum and will now cycle endlessly through the clusterings $\mathbf{z}^t, \mathbf{z}^{t+1}, \ldots, \mathbf{z}^{t'} = \mathbf{z}^t, \mathbf{z}^{t+1}, \ldots, \mathbf{z}^{t'} = \mathbf{z}^t, \ldots$.

Problem 3 (Short Answer: 4 x 8 = 32 marks). For each of the problems, give your answer in space provided.

1. We wish to perform binary classification when we have two prototypes: the triangle prototype $(0, 1)$ and the circle prototype $(1, 0)$. Find the decision boundary when we use the L_1 metric to calculate distances i.e. $d(\mathbf{z}^1, \mathbf{z}^2) = \left\|\mathbf{z}^1 - \mathbf{z}^2\right\|_1 = |\mathbf{z}_1^1 - \mathbf{z}_1^2| + |\mathbf{z}_2^1 - \mathbf{z}_2^2|$ for $\mathbf{z}^1, \mathbf{z}^2 \in \mathbb{R}^2$. Calculate the decision boundary only within the box $B := \{\mathbf{z} \in \mathbb{R}^2 : \mathbf{z}_1, \mathbf{z}_2 \in [0, 1]\} \subset \mathbb{R}^2$ and write its expression below. Draw the decision boundary in the figure. Note that you dont have to calculate the decision boundary outside the box B.

The decision boundary is the region $|x| + |y - 1| = |x - 1| + |y|$. Within the given box we have $|x - 1| = 1 - x, |x| = x, |y - 1| = 1 - y, |y| = y$. Thus, the decision boundary is $x + 1 - y = 1 - x + y$ i.e. $x = y$.



2. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable convex function on $\mathbb{R}^d$. Prove (with detailed steps) that the set $S_f := \{\mathbf{x} : f(\mathbf{x}) \leq 0\}$ is always a convex set. Use any definition of convexity you are comfortable with.

Solution: Suppose $\mathbf{x}, \mathbf{y} \in S_f$ i.e $f(\mathbf{x}) \leq 0$ and $f(\mathbf{y}) \leq 0$. Since f is convex, for any $\lambda \in [0, 1]$, if we denote $\mathbf{z} = \lambda \cdot \mathbf{x} + (1 - \lambda) \cdot \mathbf{y}$, then we have $f(\mathbf{z}) \leq \lambda \cdot f(\mathbf{x}) + (1 - \lambda) \cdot f(\mathbf{y}) \leq 0$. Thus $f(\mathbf{z}) \in S_f$ as well. This proves that S_f is convex.

3. Consider the following optimization problem for linear regression $\mathbf{x}^i \in \mathbb{R}^d, y^i \in \mathbb{R}$. In the box below, write down a likelihood distribution for $\mathbb{P}\left[y^i \,|\, \mathbf{x}^i, \mathbf{w}\right]$ and prior $\mathbb{P}\left[\mathbf{w}\right]$ such that $\hat{\mathbf{w}}_{\text{rnc}}$ is the MAP estimate for your model. Give explicit forms for the density functions but you need not calculate normalization constants.

$$\hat{\mathbf{w}}_{\text{rnc}} = \underset{\mathbf{w} \in \mathbb{R}^d}{\arg\min} \sum_{i=1}^{n} (y^i - \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle)^2 + \left\|\mathbf{w}\right\|_2^2$$
$$\text{s.t. } \left\|\mathbf{w}\right\|_2 \leq 1.$$

Solution: For any $\sigma > 0$, we can have:
Likelihood density function: $\mathbb{P}\left[y^i \,|\, \mathbf{x}^i, \mathbf{w}\right] = \mathcal{N}(\left\langle \mathbf{w}, \mathbf{x}^i \right\rangle, \sigma)$, and
Prior density function: $\mathbb{P}\left[\mathbf{w}\right] = C \cdot \mathcal{N}(\mathbf{0}, \sigma^2 \cdot I)$ if $\left\|\mathbf{w}\right\|_2 \leq 1$, else $\mathbb{P}\left[\mathbf{w}\right] = 0$,
where $C^{-1} = \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \int_{\left\|\mathbf{x}\right\|_2 \leq 1} \exp\left(-\frac{1}{2\sigma^2} \cdot \left\|\mathbf{x}\right\|_2^2\right)$

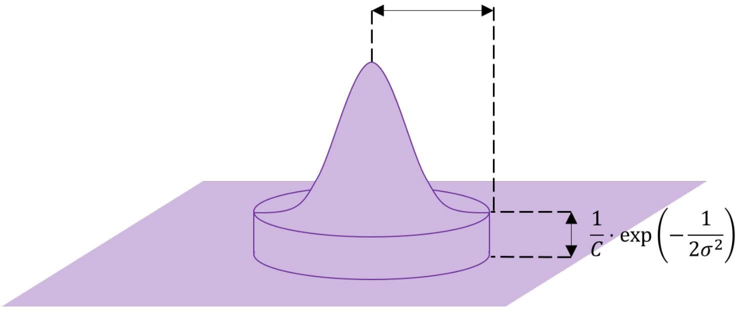


Figure 1: Depiction of the density function of the prior distribution in $d = 2$ dimensions

4. Recall that we derived the k-means algorithm by considering a Gaussian mixture model and forcibly setting the mixture proportions to $\boldsymbol{\pi}_k^t = \frac{1}{K}$ as well as the covariance matrices of the Gaussians to identity $\Sigma^{k,t} = I$. Suppose we instead set $\Sigma^{k,t} = \Sigma$ where $\Sigma \in \mathbb{R}^{d \times d}$ is a known positive definite matrix. How will the k-means algorithm change due to this? Give the final algorithm below (no derivations required).

```
Algorithm 1: tk-Means: The twisted k-Means Algorithm
Input: Number of clusters $k$, metric parameter $\Sigma$, data $\mathbf{x}^1, \ldots, \mathbf{x}^n$
$\boldsymbol{\Theta}^0 \leftarrow \{\boldsymbol{\mu}^{1,0}, \ldots, \boldsymbol{\mu}^{k,0}\}$                    //Initialize
for $t = 1, 2, \ldots,$ do
    for $i = 1, 2, \ldots, n$ do
        $z^{i,t} = \arg\min_{k \in [K]} (\mathbf{x}^i - \boldsymbol{\mu}^{k,t})^\top \Sigma^{-1} (\mathbf{x}^i - \boldsymbol{\mu}^{k,t})$
                    //Assign every data point to the "nearest" cluster center
    end for
    for $k = 1, 2, \ldots, K$ do
        $n^{k,t} = |i : z^{i,t} = k|$
        $\boldsymbol{\mu}^{i,t+1} = \frac{1}{n^{k,t}} \sum_{i : z^{i,t} = k} \mathbf{x}^i$          //Update the cluster centers
    end for
end for
```

Problem 4 (Long Answer: 3+3+5+5=16 marks). In this question we will derive an MLE estimate for a multinoulli distribution. Consider a K-faced die with faces $k = 1, 2, \ldots, K$. Let the vector $\boldsymbol{\pi}^*$ denote the vector encoding the probabilities of the various faces turning up i.e. face k turns up with probability $\boldsymbol{\pi}_k^*$. Clearly $\boldsymbol{\pi}_k^* \geq 0$ and $\sum_{k=1}^K \boldsymbol{\pi}_k^* = 1$. Now suppose I get n rolls of this die. Let $\mathbf{x} \in \mathbb{N}^K$ denote the vector that tells me how many times each face turned up i.e. the k-th face is found turning up $\mathbf{x}_k \geq 0$ times with $\sum_{k=1}^K \mathbf{x}_k = n$ (recall $\mathbb{N} = \{0, 1, 2, \ldots\}$ is the set of natural numbers). It turns out that we have $\mathbb{P}[\mathbf{x} \,|\, \boldsymbol{\pi}^*] = \frac{n!}{\prod_{k=1}^K (\mathbf{x}_k!)} \prod_{k=1}^K (\boldsymbol{\pi}_k^*)^{\mathbf{x}_k}$.

1. Write down the problem of finding the MLE estimate $\arg\max_{\boldsymbol{\pi}} \mathbb{P}[\mathbf{x} \,|\, \boldsymbol{\pi}]$ as an optimization problem. *Hint*: it will be a constrained optimization problem.

Solution: The following optimization problem captures the task of learning a discrete probability distribution which should have positive mass on its support as well as well as should be normalized

$$\min_{\boldsymbol{\pi} \in \mathbb{R}^K} \; -\sum_{k=1}^K \mathbf{x}_k \log \boldsymbol{\pi}_k$$

$$\boldsymbol{\pi}_k \geq 0, \text{ for all } k \in [K]$$

$$\sum_{k=1}^K \boldsymbol{\pi}_k = 1$$

The positivity constraint is not super essential since we can ask logarithm function to itself rule out solutions that do not have $\boldsymbol{\pi}_k \geq 0$ by setting $\log v = -\infty$ for all $v \leq 0$. However, if we wish to be careful, we should put a positivity constraint in. We will see that such positivity constraints never really bother us while preparing the dual problems or solving them.

2. Write down the Lagrangian for that optimization problem.

Solution: There are $K+1$ constraints here so there will be those many Lagrange multiplier variables. We will have a vector Lagrange variable $\boldsymbol{\alpha} \in \mathbb{R}^K$ to account for the positivity constraints and one more variable λ to account for the normalization constraint.

$$\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\alpha}, \lambda) = -\sum_{k=1}^K \mathbf{x}_k \log \boldsymbol{\pi}_k - \sum_{k=1}^K \boldsymbol{\alpha}_k \cdot \boldsymbol{\pi}_k + \lambda \cdot (\sum_{k=1}^K \boldsymbol{\pi}_k - 1)$$

Notice the choice of negative sign in $-\alpha_k \boldsymbol{\pi}_k$. This is because the inequality we have is of the form $\boldsymbol{\pi}_k \geq 0$ which is opposite of the canonical form.

3. Find the dual problem and eliminate the primal variable. Show major steps. Give the simplified dual problem which should be only in terms of constants and the dual variable.

Solution: Verify for yourself that the primal problem

$$\min_{\boldsymbol{\pi}} \max_{\boldsymbol{\alpha} \geq 0, \lambda} \; \mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\alpha}, \lambda)$$

does indeed solve the original problem in part 1 of this question. The dual problem is

$$\max_{\boldsymbol{\alpha} \geq 0, \lambda} \min_{\boldsymbol{\pi}} \; \mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\alpha}, \lambda)$$

Recall that $\boldsymbol{\alpha} \geq 0$ is notation to denote that all coordinates of the vector are constrained to be non-negative. Notice that since $\boldsymbol{\alpha}_k$ correspond to inequality constraints, they must be forced to take positive values to ensure that we do not change the optimization problem while creating the dual. Also notice that since λ corresponds to an equality constraint, no constraints are imposed on it (one can impose $\lambda \neq 0$ but it is unnecessary).

Since the inner optimization problem is unconstrained, we apply first order optimality condition to get the following

$$-\frac{\mathbf{x}_k}{\boldsymbol{\pi}_k} - \boldsymbol{\alpha}_k + \lambda = 0, \text{ for all } k \in [K],$$

which allows us to eliminate the primal variable as $\boldsymbol{\pi}_k = \frac{\mathbf{x}_k}{\lambda - \boldsymbol{\alpha}_k}$, to get the simplified dual problem

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^K, \lambda \in \mathbb{R}} \; -\sum_{k=1}^K \mathbf{x}_k \log\left(\frac{\mathbf{x}_k}{\lambda - \boldsymbol{\alpha}_k}\right) - \sum_{k=1}^K \frac{\boldsymbol{\alpha}_k \cdot \mathbf{x}_k}{\lambda - \boldsymbol{\alpha}_k} + \lambda \cdot \left(\sum_{k=1}^K \frac{\mathbf{x}_k}{\lambda - \boldsymbol{\alpha}_k} - 1\right)$$
$$\boldsymbol{\alpha}_k \geq 0, \text{ for all } k \in [K]$$

The objective function further simplifies to give us

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^K, \lambda \in \mathbb{R}} \; -\sum_{k=1}^K \mathbf{x}_k \log\left(\frac{\mathbf{x}_k}{\lambda - \boldsymbol{\alpha}_k}\right) + \sum_{k=1}^K \mathbf{x}_k - \lambda$$
$$\boldsymbol{\alpha}_k \geq 0, \text{ for all } k \in [K]$$

4. Solve the dual problem and use it to obtain the MLE estimate. Only give expressions for both the dual solution as well as the MLE estimate.

Solution: First of all we notice that the objective function decreases with increasing value of $\boldsymbol{\alpha}_k$ for every $k \in [K]$. Since we have a maximization problem at hand, this forces us to set $\boldsymbol{\alpha}_k$ to its lowest allowed value i.e. 0 for all $k \in [K]$. Notice that the complimentary slackness condition also suggests something similar since it tells us that $\boldsymbol{\alpha}_k \mathbf{x}_k = 0$.

However, be careful to note that the complimentary slackness condition has not been used here to solve the dual problem. We independently arrived at the conclusion that $\boldsymbol{\alpha}_k$ must be set to zero (using the KKT conditions to simplify the dual problem gives something called the Wolfe dual problem which we are not doing here). Setting $\boldsymbol{\alpha}_k = 0$ gives us the simplified problem

$$\max_{\lambda \in \mathbb{R}} \sum_{k=1}^{K} \mathbf{x}_k \log\left(\frac{\mathbf{x}_k}{\lambda}\right) + \sum_{k=1}^{K} \mathbf{x}_k - \lambda$$

Notice that in the problem we derived above, the optimization over λ is unconstrained. Applying the first order optimality condition over λ (i.e. setting the partial derivative of the objective with respect to λ to zero) gives us

$$\lambda = \sum_{k=1}^{K} \mathbf{x}_k$$

Recall that we had

$$-\frac{\mathbf{x}_k}{\boldsymbol{\pi}_k} - \boldsymbol{\alpha}_k + \lambda = 0$$

We use this, and the optimal solutions to $\boldsymbol{\alpha}, \lambda$ obtained above to get

$$\boldsymbol{\pi}_k = \frac{\mathbf{x}_k}{\sum_{k=1}^{K} \mathbf{x}_k}$$