End sem grading rubric

T/F answers (for reference)

1. N
2. 10110
3. $N(w|w_0, \sigma^2 I)$
4. O(D), O(N)
5. 1NN< DT < LR < kSVM, DT < LR < kSVM < 1NN
6. $\pi_k = \frac{1}{K} \forall k \in \{1 \cdots K\}$, $\Sigma_k = \sigma^2 \mathcal{I} \forall k \in \{1 \cdots K\}$
7. $0.5Bern(x|\pi_1) + 0.5Bern(x|\pi_2)$
8. $O(D^2)$, ∞

For Qs 5 and 6, give 2 marks for each correct blank filled. For all questions, if answers are not correct, but show some glimmers of understanding, give half marks.

Parameter estimation

(a) Poisson rate parameter MLE

- give +1 mark for correctly calculating the likelihood function given n data samples $x_1, \cdots x_n$ as a product of n Poissons
- give +2 mark for correctly reducing to log form
- give +2 mark for correctly calculating the derivative w.r.t. λ
- give +1 mark for getting to the right answer, viz. the MLE for λ is the sample mean $\frac{1}{n}\sum_{i=1}^{n} x_i$

(b) Dating disasters

- give +2 marks for correctly realizing that the likelihood is a product of Bernoulli trials $p^2(1-p)^6$ and the prior is Beta(2,2), which effectively reduces to $p(1-p)$
- give +2 marks for correctly calculating the MLE using the likelihood by first taking the log and then differentiating to get 0.25. Give zero for only writing the answer.
- give +2 marks for correctly calculating the MAP using the posterior by multiplying the likelihood with the prior, then taking the log and then differentiating to get 0.3. Give zero for only writing the answer.
- give +2 marks for correctly specifying that the prior is updated to Beta(4,8).

(c) MLE anomaly

- give +1 point for correctly specifying the log likelihood as $L(\mu) = -\frac{1}{2}(x-\mu)^2 - \frac{1}{2}\log(2\pi)$
- give +1 point for calculating MLE for μ as x.
- give +1 point for realizing that this can only remain true when $x \in [a, b]$
- give +1 point for any progress at all towards realizing that the MLE calculated analytically is incomplete.

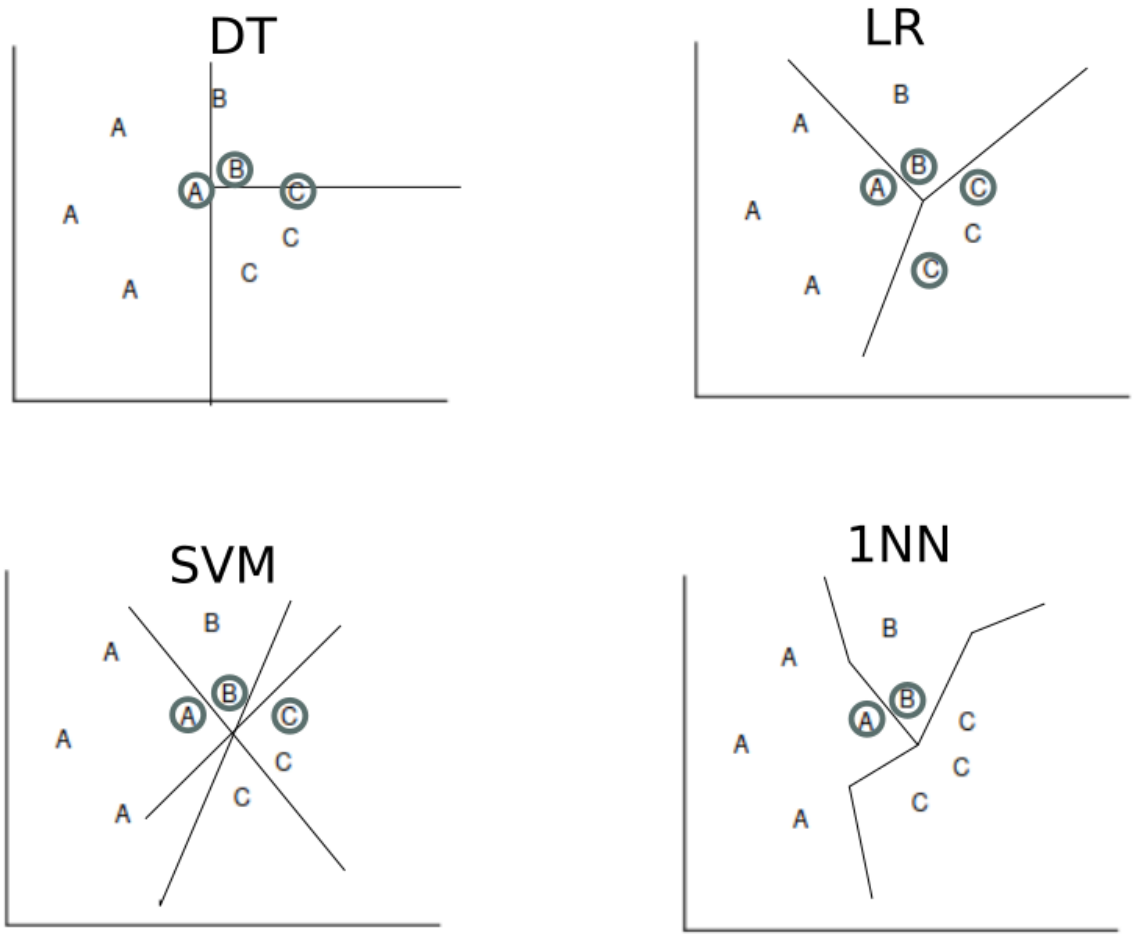
- give +2 points for getting to the right answer that $\mu = x$ when x is between a and b, is equal to a when $x < a$ and b when $x > b$

Classic classifiers

(a) Tweaks to classifiers

- Give +1 for mentioning the need to discretize real x,y coordinates for DT to work
- Give +1 for mentioning multinomial logistic regression
- Give +1 for mentioning either pairwise or one-vs-all for SVM
- Give +1 for saying the 1-NN will work fine without changes

(b)-(c) for both these parts, look at the image below and mark answers using the associated rubric.



For (b)

- Give +1 for getting the DT boundaries right
- Give +1 for getting the LR boundaries right
- Give +2 for getting the SVM boundaries right (I am showing boundaries for pairwise training, one vs all will be different. make sure peoples' boundaries match what they wrote in (a))
- Give +2 for getting 1NN boundaries right

For (c)

- Give -1 for each point missed in any of the four graphics (obviously capping the cuts at 6 points)

For (d)

- Give +2 for reporting the multi-class confusion matrix itself or the multi-class F1 score in a one-vs-all comparison.
- Give +1 for pointing out that the F score ignores true negatives
- Give +1 for pointing out that one-vs-rest F score is spuriously inflated in settings with few positive and many negative examples because it ignores true negatives

SVMs and optimization

(a) Finding the SVM margin

- give +1 point for saying that the SVM decision boundary tries to maximize the margin between the two classes
- give +1 point for realizing that the points halfway between the support vectors on both sides are (3,4) and (2,5) and realizing that this means the decision boundary is x+y = 7, so $w_1 = w_2$
- give +1 point for plugging support vector coordinates into the hyperplane equation to get an equation pair like $2w_1 + 3w_2 + b = 1$ and $4w_1 + 5w_2 + b = -1$
- give +1 point for solving the system of three equations for $\{w_1, w_2, b\}$. I get them as $\{-0.5, -0.5, 3.5\}$
- give +1 point for saying there are 5 support vectors in the dataset
- give +1 point for calculating the margin as $2/\|w\| = 2\sqrt(2)$

(b) Proving NN square loss non-convexity. Here is one possible approach.

- Consider a path in a simple 3 layer NN, where an input node connects to a hidden node through weight w_2 and the hidden node connects to the output through weight w_1.
- Give input x and output y, squared error loss here would be $L(\mathbf{w}) = (y - w_2 w_1 x)^2$. Give +3 for being able to state this much.
- We have to show that $L(\mathbf{w})$ is non-convex in $\mathbf{w}$. We can do this by showing violations of Jensen's inequality for some values of $\mathbf{w}$. Give +2 for proposing a coherent strategy like this one for the proof.
- Give +2 for applying Jensen's inequality (or whatever other result one is using to prove non-convexity) correctly.
- Give +3 for successfully demonstrating a counter-example. I give one such demonstration below.

Assume $x = 1, \mathbf{w} = [2, 1/2]^T$ and some $\mathbf{u} = [-1, 1]^T$ so that if $f(\mathbf{w}; x) = w_2 w_1 x$ is convex, by Jensen's inequality, for any non-negative α,

$$f(\alpha\mathbf{w} + (1-\alpha)\mathbf{u}) \leq \alpha f(\mathbf{w}) + (1-\alpha) f(\mathbf{u}).$$

Further, define $\mathbf{v} = \alpha\mathbf{w} + (1-\alpha)\mathbf{u}$, which equals $[3\alpha - 1, 1 - \alpha/2]^T$ for our choice of $\mathbf{w}, \mathbf{u}$.

Plugging back into the Jensen's inequality, the LHS becomes $f(\mathbf{v}) = (3\alpha - 1)(1 - \alpha/2)$, while the RHS becomes $\alpha(1) + (1 - \alpha)(-1) = 2\alpha - 1$. It can be easily shown that the LHS is only less than the RHS if $\alpha > 1$, so for all $\alpha \in [0, 1)$, we have shown a contradiction.

Thus the function is non-convex.

(c) Gradient descent variations

- Give +1 for correctly identifying the difference between GD and SGD, viz. one point used for gradient calculation in SGD vs the whole dataset used in GD
- Give +1 for correctly identifying computational savings as the advantage of SGD over GD
- Give +1 for correctly identifying the difference between mini-batch SGD and SGD, viz. using a small batch of points instead of just one point for gradient calculations
- Give +1 for correctly identifying stability of the gradient calculation as the advantage of minibatch SGD over SGD

EM and clustering

(a) Mixture of exponentials EM design

- Give +1 for saying that we will run the E-step and M-step alternately until convergence
- Give +3 for calculating the responsibility correctly in the E-step

$$z_{nk} = \frac{\pi_k \lambda_k e^{-\lambda_k x_n}}{\sum_{k=1}^{K} \pi_k \lambda_k e^{-\lambda_k x_n}}$$

- Give +3 for deriving the correct log likelihood to minimize in the M step

$$L = \sum_{nk} z_{nk}(\lambda_k x_n - \log \lambda_k - \log \pi_k)$$

- Give +2 for adding a Lagrangian to handle the mixture weights normalization constraint

$$L = \sum_{nk} z_{nk}(\lambda_k x_n - \log \lambda_k - \log \pi_k) + \gamma(\sum_{k} \pi_k - 1)$$

- Give +3 for differentiating and calculating each of the parameter updates correctly. The correct updates are given below.

$$\pi_k = \frac{\sum_n z_{nk}}{\sum_{nk} z_{nk}}$$

$$\gamma = \sum_{nk} z_{nk}$$

$$\lambda_k^{-1} = \frac{\sum_n z_{nk} x_n}{\sum_n z_{nk}}$$

(b) k-means to PCA

- Give +1 for saying that z_n goes from a one-hot vector to a real vector.

- Give +2 for saying that we add orthonormality constraints $\mathbf{w}_i^T \mathbf{w}_j = 0$ for $i \neq j$ and $\|\mathbf{w}_i\|^2 = 1$ to the optimization.

(c) Reconstructing matrix from eigenvalues and eigenvectors

- Give +1 for stating that eigendecomposition factorizes a matrix $M = ABA^{-1}$, where A is a matrix made up of the eigenvectors, and B is a diagonal matrix containing the eigenvalues.
- Give +1 for correctly creating A = (-1 1; 2 1) and B = (-1 0; 0 3).
- Give +1 for correctly inverting A
- Give +2 for putting the pieces together and calculating M = (5/3 4/3; 8/3 1/3). Being off by a scale factor makes no difference. Half marks if people don't complete the calculation.