Machine Learning Midterm Examination

University of Texas at Dallas

10/24/2022

Question	Topic	**Points**
1	Short Answers	20
2	Duality	15
3	Probabilistic Models	20
4	Support Vector Machines	20
5	Linear Regression	13
6	Decision Trees	12
Total		100

Instructions:

1. You have **two and a half (2.5) hours** to complete the examination.
2. Please show all the steps clearly of how you came up with the final answer. Just the final answer with no work will be zero points!!
3. Either you can use this paper or a separate set of sheets to fill in your answers. Write clearly so we can understand your handwriting.
4. Please order your questions according to the questions and write in clear handwriting so it is easy for us to grade. Otherwise we will deduct 10 points.
5. Please do not search online for answers to the questions. If the answers are similar to something available online, you will get zero points on this examination.
6. The examination has to be done individually by everyone. If someone copies, the entire group of students involved will get a zero.
7. Work efficiently. Some questions are easier, some more difficult. Be sure to give yourself time to answer all of the easy ones, and avoid getting bogged down in the more difficult ones.
8. All the Best!!

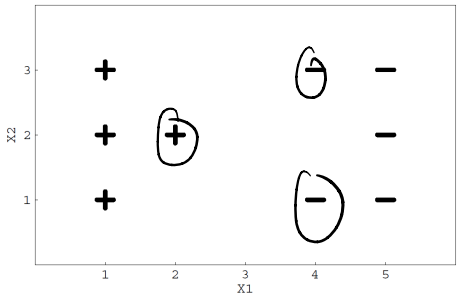
Question 1: Short Answers

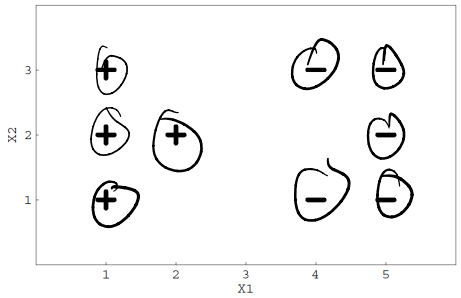
[20 pts] Please provide short and clear answers for the questions below. Please explain your answer and show your work. No credit if the explanation is incorrect.

(a) (6 points) Consider the 2 dimensional dataset given below. Circle examples having the following property: removing any of the examples and retraining the classifier would yield a different decision boundary than training on the full dataset. Circle examples for the following classifiers:

- Left Figure: Linear Support Vector Machines
- Right Figure: Logistic Regression

Briefly explain why.



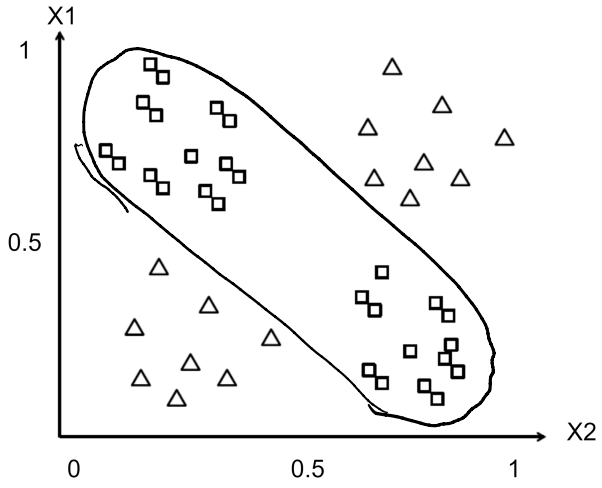


SVM: Only support vectors matter

LR: All points matter for DB

(b) (5 points) This question has three parts. Refer to the dataset below (triangles are negatives and squares are positives).

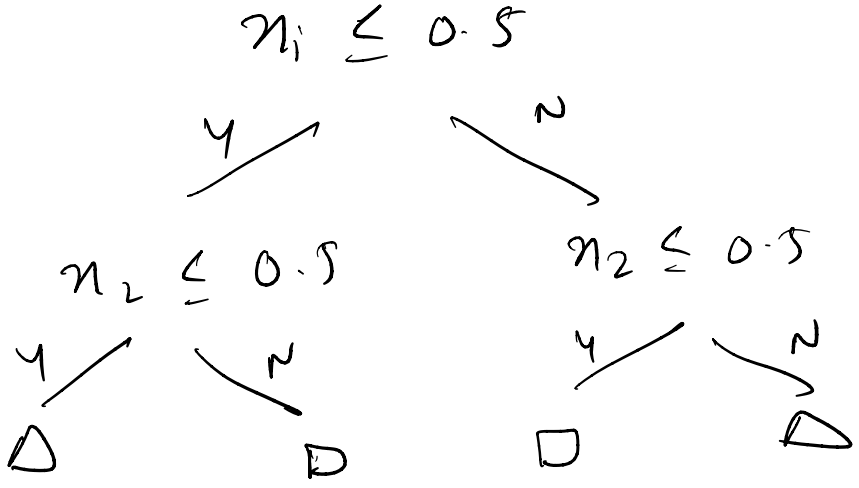
- (1 point) Can this dataset be perfectly classified (i.e. zero training error) with a linear classifier?
- (2 points) Give an example of a kernel such that a kernel SVM will have zero training error.
- (2 points) Draw a decision tree having zero training error.



1. No

2. Gaussian kernel or Poly kernel, degree 2

3.



(c) (3 points) For linearly separable data, can a small slack penalty ("C") hurt the training accuracy when using a linear SVM (with no kernel)? If so, explain how and if not, why not?

Yes. If C is too small, it can cause over-regularization.

$C \to 0 \Rightarrow$ Arbitrary Classifier.

(d) (3 points) Imagine we are running stochastic gradient descent, and we determine the stopping criteria based on the validation set. Suppose the stopping criteria is if the validation set error increases (i.e. stop SGD as soon as the validation set error starts increasing). Is this a good stopping criteria? If not, how would you fix it?

It is a good stopping criteria for well-behaved losses. For non-convex losses, sometimes val error increases then decreases.

Fix: Have a patience param to wait for few epochs

(e) (3 points) Construct a one-dimensional classification dataset for which the Leave-one-out cross-validation error of the One Nearest Neighbors algorithm is always 1 (i.e. 100% error). Stated another way, the

+ − + − + − + −

LOOCV error = 100%.

One Nearest Neighbor algorithm never correctly predicts the held out point.

Question 2: Duality

[15 pts] This question is on computing the dual of certain loss functions and the relationship between the primal and the dual problem.

(a) Part 1 (7 Points) Ridge Regression: Ridge Regression is L2-Regularized Least Squares. Recall that, given training data $\{(x^1, y^1), \cdots, (x^M, y^M)\}$, the optimization problem is:

$$\min_w \sum_{i=1}^{M} (y^i - w^T x^i)^2 + \lambda ||w||^2 \tag{1}$$

Here we assume the bias is accounted for in the weight vector. For simplicity, we can also write it in Matrix form:

$$min_w ||Y - Xw||^2 + \lambda ||w||^2 \tag{2}$$

where Y is the column vector $[y^1, \cdots, y^M]^T$ and $X \in \mathbb{R}^{M \times d}$ is a Matrix such that ith row is x^i (for $i = 1, \cdots, M$). Compute the dual of this expression. Hint: Introduce slack variables $\zeta^i = y^i - w^T x^i$, and then write down an optimization problem over w and ζ. Note that to solve this problem, you should have either equality or inequality constraints.

$$\min_{w} \quad ||\Gamma||^2 + \lambda ||w||^2$$

$$s.t \quad \Gamma = Y - Xw$$

$$L(w, \Gamma, \alpha) = ||\Gamma||^2 + \lambda ||w||^2 + \alpha (Y - Xw - \Gamma)$$

$$\frac{\partial L}{\partial \Gamma_i} = 2\Gamma - \alpha = 0$$

$$\Rightarrow \Gamma_i = \alpha_i / 2 \quad \text{---------- (1)}$$

$$\frac{\partial L}{\partial w} = 2\lambda w - \sum_{i=1}^{M} \alpha_i x_i = 0$$

$$\Rightarrow w = \frac{\sum_{i=1}^{M} \alpha_i x_i}{2\lambda} \quad \text{— (2)}$$

$$L(\alpha) = \sum_{i=1}^{M} \frac{\alpha_i^2}{4} + \lambda \frac{\left\| \sum_{i=1}^{M} \alpha_i x_i \right\|^2}{4\lambda^2}$$

$$+ \sum_{i=1}^{M} \alpha_i \left[y_i - \frac{\sum_{j=1}^{M} \alpha_j x_i^T x_j}{2\lambda} - \frac{\alpha_i}{2} \right]$$

Find Exp

$$L(\alpha) = \sum_{i=1}^{M} -\frac{\alpha_i^2}{4} - \frac{1}{4\lambda} \sum_{(i,j)} \alpha_i \alpha_j x_i^T x_j + \sum_i \alpha_i y_i$$

(b) Part 2 (5 Points) Consider the L2 Regularized Square SVM:

$$\min_{w} \sum_{i=1}^{M} [\max(0, 1 - y^i(w^T x^i + b))]^2 + \lambda ||w||^2 \tag{3}$$

Compute the dual of this objective function. Same hint as above. I.e., introduce slack variables $\zeta^i = 1 - y^i(w^T x^i + b)$ or possibly $\zeta^i = \max(1 - y^i(w^T x^i + b), 0)$, and then write down an optimization problem over w, b and ζ.

$$\min_{w,b,\zeta_i} \sum_{i=1}^{M} \zeta_i^2 + \lambda \|w\|^2$$

$$\text{s.t} \quad y_i(w^T x_i + b) \geq 1 - \zeta_i$$

$$\zeta_i \geq 0$$

$$L(\alpha, \beta, w, b, \zeta) = \sum_{i=1}^{M} \zeta_i^2 + \lambda \|w\|^2 + \sum_{i=1}^{M} \alpha_i [1 - \zeta_i - y_i(w^T x_i + b)] - \beta_i \zeta_i \quad \text{— (1)}$$

$$\frac{\partial L}{\partial \zeta_i} = 2\zeta_i - \alpha_i - \beta_i$$

$$\Rightarrow \zeta_i = \frac{\alpha_i + \beta_i}{2} \quad \text{— (2)}$$

$$\frac{\partial L}{\partial w} = 2\lambda w - \sum_{i=1}^{M} y_i x_i \alpha_i = 0$$

$$\Rightarrow w = \sum_{i=1}^{M} \frac{y_i x_i \alpha_i}{2\lambda} \quad \text{— (3)}$$

$$\frac{\partial L}{\partial b} = \sum_i \alpha_i y_i = 0 \quad \text{— (4)}$$

Substitute (2) & (3) into (1) with

Constraints: (4) & $\alpha_i \geq 0, \beta_i \geq 0$

Find Expression

$$\max_{\alpha_i, \beta_i} \; -\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$-\frac{1}{2}(\alpha_i + \beta_i)^2 + 4\sum_{i=1}^{m} \alpha_i$$

$$s.t \quad \sum_{i=1}^{m} \alpha_i y_i = 0 \;, \; \alpha_i \geq 0 \;, \; \beta_i \geq 0$$

(c) Recall that the Kernel SVM is:

$$\max_{\lambda \geq 0, \sum_i \lambda^i y^i = 0} -1/2 \sum_i \sum_j \lambda^i \lambda^j y^i y^j K(x^i, x^j) + \sum_i \lambda^i \tag{4}$$

Given a solution λ^* of the dual expression above, and given a new test point x^t, how will you obtain the prediction for x^t? Give an example of the prediction if the Kernel is $K(x^i, x^j) = \exp(||x^i - x^j||^2/2\sigma^2)$.

First compute b using complementary slackness

$$b = y_i - w^T \phi(x_i)$$

$$= y_i - \sum_j \lambda_j y_j \phi(x_j)^T \phi(x_i)$$

$$= y_i - \sum_j \lambda_j y_j K(x_i, x_j)$$

Next, $y_t = w^T \phi(x_t) + b$

$$= \sum_j y_j \lambda_j K(x_i, x_t) + b$$

For Gauss kernel, substitute

$$K(x, y) = \exp\left(-\|x-y\|^2 / 2\sigma^2\right)$$

Question 3: Probabilistic Models and Maximum Likelihood Estimation

[20 pts] This question consists of three parts.

Consider a positive, real-valued random variable X that is distributed according to a log-normal distribution:

$$p(x) \propto \frac{1}{x\sigma\sqrt{2\pi}} \exp(-(\log x - \mu)^2/\sigma^2) \tag{5}$$

for real values parameters μ and $\sigma > 0$. Suppose you are given M data points $x^1, \cdots, x^M$.

- (2 points) Compute the Log Likelihood of the data observations.
- (8 points) Find the MLE estimators for μ and σ. Are these unbiased?
- (10 points) Find the MAP estimate for μ if the the prior distribution of μ is itself distributed as a Normal distribution with mean ν and variance β^2. First compute the Posterior distribution and then maximize the parameters to obtain the mean.

(a) $\prod_{i=1}^{M} p(x^{(i)}) = \prod_{i=1}^{M} \frac{1}{x^i \sigma\sqrt{2\pi}} \exp\left(-\frac{(\log x_i - \mu)^2}{\sigma^2}\right)$

$$LL = \sum_{i=1}^{M} -\frac{(\log x_i - \mu)^2}{\sigma^2} - M \log x^i \sigma\sqrt{2\pi}$$

(b) $\frac{\partial LL}{\partial \mu} = \sum_{i=1}^{M} \frac{2\log x_i}{\sigma^2} - \frac{2\mu M}{\sigma^2} = 0$

$$\Rightarrow \mu = \frac{\sum_{i=1}^{M} \log x_i}{M}$$

$$\frac{\partial L}{\partial \sigma} = \frac{\sum_{i=1}^{m} 2(\log x_i - \mu)^2}{\sigma^3} - \frac{m}{\sigma} = 0$$

$$\Rightarrow \sigma = \sqrt{\frac{2\sum_{i=1}^{m} (\log x_i - \mu)^2}{m}}$$

Both μ & σ are biased since

$$E[\mu] \neq \mu \text{ & } E[\sigma^2] \neq \sigma^2$$

c) MAP: $\sum_{i=1}^{m} -\frac{(\log x_i - \mu)^2}{\sigma^2} - m\log x_i \sigma\sqrt{2\pi} + \frac{(\mu - v)^2}{B^2}$

$$\frac{\partial MAP}{\partial \mu} = \frac{2\sum_{i=1}^{m} \log x_i}{\sigma^2} + \frac{2m\mu}{\sigma^2} + \frac{2(\mu - v)}{B^2}$$

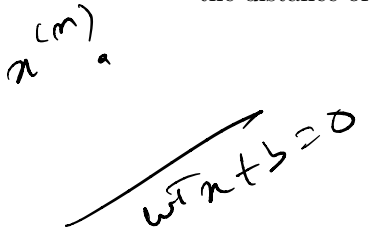
$$\Rightarrow \mu\left(\frac{M}{\sigma^2} + \frac{1}{\beta^2}\right) = \frac{\sum_{i=1}^{m} \log x_i}{\sigma^2} + \frac{\nu}{\beta^2}$$

$$\mu_{MAP} = \frac{\frac{\sum_{i=1}^{M} \log x_i}{\sigma^2} + \frac{\nu}{\beta^2}}{\frac{M}{\sigma^2} + \frac{1}{\beta^2}}$$

Question 4: Support Vector Machines

[20 pts] Consider a binary classification problem for vectors in $\mathbb{R}^n$ using linear separators. For this problem, consider linearly separable, labeled training data of the form $(x^{(1)}, y^{(1)}), \ldots, (x^{(M)}, y^{(M)})$, where $x^{(m)} \in \mathbb{R}^n$ and $y^{(m)} \in \{+1, -1\}$.

(a) (4 points) Given a max-margin linear separator of the form $w^T x + b$ for the above datapoints, what is the distance of the point $x^{(m)}$ to the linear separator as a function of w and b?



Distance of $x^{(m)}$ to $w^T x + b = 0$ is $\frac{|w^T x^{(m)} + b|}{\|w\|}$

(b) (4 points) Express the constraint that "$x^{(m)}$ cannot be farther away from the linear separator than δ times the size of the margin" as a pair of linear inequalities, for some constant $\delta > 0$.

$$\frac{|w^T x^{(m)} + b|}{\|w\|} \leq \frac{\delta}{\|w\|}$$

(c) (8 points) Add the constraints from part (b) for each $x^{(m)}$ to the standard SVM objective without

slack. Construct a dual of this optimization problem, treating δ as a hyperparameter, using the method of Lagrange multipliers.

svm + (b) constraints:

$$\min_{w,b} \quad \|w\|^2$$

s.t: $y_i(w^T x_i + b) \geq 1, \quad \forall i$

$$|w^T x_i + b| \leq \delta, \quad \forall i$$

$$w^T x + b \leq \delta \quad \& \quad -(w^T x_i + b) \leq \delta$$

$$\mathcal{L}(w, b, \alpha, \beta, \gamma) = \|w\|^2 + \sum_{i=1}^{m} \alpha_i \left[1 - y_i(w^T x_i + b)\right] + \sum_{i=1}^{m} \beta_i \left(w^T x_i + b - \delta\right] - \sum_{i=1}^{m} \gamma_i (w^T x_i + b + \delta]$$

(*)

$$\frac{\partial \mathcal{L}}{\partial w} = 2w - \sum_{i=1}^{m} \alpha_i y_i x_i + \sum_{i=1}^{m} (\beta_i - \gamma_i) x_i = 0$$

$$\Rightarrow w = \frac{\sum_{i=1}^{m} \left[\alpha_i y_i x_i - (\beta_i - \gamma_i) x_i\right]}{2} \quad \text{— (1)}$$

$$-\sum_{i=1}^{m}\left[\alpha_i y_i - (\beta_i - \gamma_i)\right] = 0 \quad \cdots \frac{\partial L}{\partial b} = 0$$

— (2)

substitute (1) back in (*) & use (2) as constraints

(d) (4 points) Explain the effect δ has on the solution of the modified SVM. How would you pick δ in practice?

δ is a hyper-parameter that tries to have all datapoints close to the DB on both sides.

Treat δ as a hyperparameter & use cross-validation or val set to solve!

Question 5: Linear Regression

[13 pts] Consider a regression task that fits a piece wise linear function of the form: $f(x) = a_1 x + b_1, x < 0$ and $f(x) = a_2 x + b_2$ if $x \geq 0$.

(a) Part 1 (5 points): Given data points $(x^{(1)}, y^{(1)}), \ldots, (x^{(M)}, y^{(M)})$, where $x^{(m)} \in \mathbb{R}^n$ and $y^{(m)} \in \mathbb{R}$, formulate a regression problem to predict y as a loss minimization problem and explain how to use gradient descent.

Let $P \subseteq [1, \ldots, M]$ s.t $x^{(i)} \geq 0$

$N \subseteq [1, \ldots, M]$ s.t $x^{(i)} < 0$

$$L(a,b) = \sum_{i \in P} [a_1 x_i + b_1 - y_i]^2 + \sum_{i \in N} [a_2 x_i + b_2 - y_i]^2$$

$$L(a,b) = L(a_1, b_1) + L(a_2, b_2)$$

$$\therefore \min_{a,b} L(a,b) = \min_{a_1, b_1} L(a_1, b_1) \; \& \; \min_{a_2, b_2} L(a_2, b_2)$$

i.e. optimize a_1, b_1, a_2, b_2 independently via gradient descent.

(b) Part 2 (8 points) If we want to apply an additional constraint that f must be continuous, then formulate regression under this new constraint as a convex optimization problem. Then, write down the Lagrangian. Can you compute the Dual expression?

$$\min_{a_1, b_1, a_2, b_2} \sum_{i \in P} (a_1 x_i + b_1 - y_i)^2 + \sum_{i \in N} (a_2 x_i + b_2 - y_i)^2$$

$$s.t \quad b_1 = b_2$$

$$\mathcal{L}(a, b, \lambda) = \sum_{i \in P} (a_1 x_i + b_1 - y_i)^2 + \sum_{i \in N} (a_2 x_i + b_2 - y_i)^2 + \lambda(b_1 - b_2)$$

You can compute dual using Matrix formulation.

$$\|A_P^1 x_P + b_1 1_P - y_P\|^2 + \|A_N^2 x_N + b_2 1_N - y_N\|^2$$

$$+ \lambda(b_1 - b_2) \Rightarrow L(A, b, \lambda)$$

where x_P = vector of +ive x_i

x_N = vector of -ive x_i

[Points will be given even if you derive the Lagrangian]

Question 6: Decision Trees

[12 pts] In class, we studied the decision tree algorithm for classification in detail. Here, you will need to derive the decision tree algorithm for regression. Assume you are given a dataset $\{(x^1, y^1), \cdots, (x^M, y^M)\}$, where the labels are continuous real numbers $y \in \mathbb{R}$.

- (5 points) First, write down the splitting condition. How will you choose the feature and/or threshold to split at every step of the recursive tree building algorithm?
- (2 points) Given the above, write down the complete decision tree algorithm.
- (4 points) Create a simple 1-D dataset with 10 instances (features and labels) and create a rough decision tree up to depth 2. I'm not expecting the exact tree but just a rough approximation.
- (1 point) What are the key hyper-parameters in the decision tree algorithm for regression?

(a) Consider split (n_s, t) which splits dataset into $x_i, i \in L_{s,t}$ & $x_j, j \in R_{s,t}$

$$\min_{s,t} \sum_{i \in L_{s,t}} \left[y_i - \frac{\sum_{j \in L_{s,t}} y_j}{|L_{s,t}|} \right]^2 + \sum_{i \in R_{s,t}} \left[y_i - \frac{\sum_{j \in R_{s,t}} y_j}{|R_{s,t}|} \right]^2$$

i.e use Least Square measure

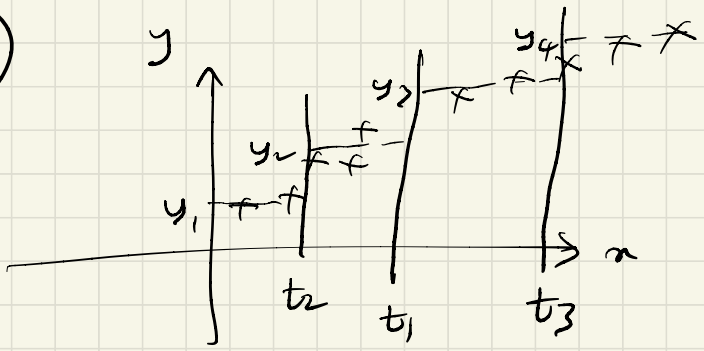
(b) Recurse over nodes & split based on

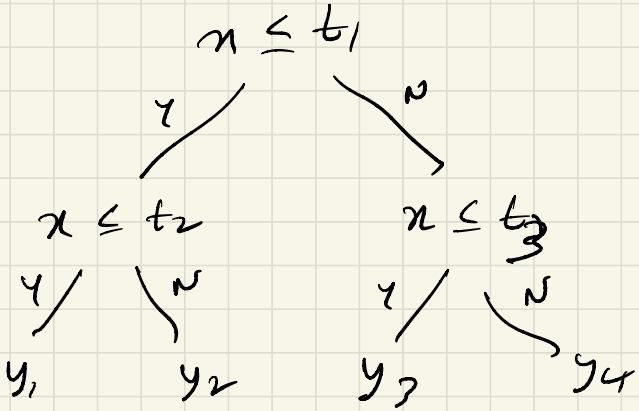
(a)

(1) At node i, find split based on (a)

(2) Recurse over left & right sides until stopping condition is met.

(c)





(d) Depth

leaf nodes

Max # instances per leaf node.