

Midterm Outline (For reference, subject to change)

L1	KNN	6%
L2	Perceptron	6%
L3	MLE, MAP	10%
L4	Naive Bayes	10%
L5	Gaussian Naive Bayes	6%
L6	Logistic Regression	6%
L7	Linear Regression	10%
L8, 9	SVM	15%
L10	Kernel	6%
L11	Theory, Decision Tree	5% + 5%
L12	Boosting, Bagging	5%
L13	Neural Nets	10%


- Roughly 35 - 40 questions
- Multiple answers may apply
- Zoom during lecture time
- Open book (lecture notes allowed)
- Do not try to use ChatGPT-like AI tools, not helpful most of the time

SVM



- Understand the geometric illustrations of SVM and relevant concepts
- Derive the SVM problem from its geometric explanation
- Understand the connection between the primal problem and the dual problem
- Understand the difference between soft-margin SVM and hard-margin SVM
- Apply Lagrangian multiplier to similar problems


Sample Question



For $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{y} \in \mathbb{R}^d$, which of the following $\kappa(\mathbf{x}, \mathbf{y})$ is **NOT** a valid kernel function?

- i. $-\mathbf{x}^\top \mathbf{y}$
- ii. $\exp(\frac{\|\mathbf{x} - \mathbf{y}\|_2^2}{2\sigma^2}) + \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|_2^2)$
- iii. $\exp(-\gamma \|\mathbf{x} - \mathbf{y}\|_2^2)$
- iv. $\tanh(1 + \mathbf{x}^\top \mathbf{y})$
- v. All of them are valid kernel functions

Sample Question



Suppose we have some data points in 2d space and we want to use a linear model to fit the data points, the loss functions we have are: (1) L1 loss; (2) L2 loss; (3) L2 loss with regularization.

Which one will more likely train a model that is closer to the outliers?


A. L1 loss.

B. L2 loss.

C. L2 loss with regularization.

D. All of them will be equally close to the outliers.

Sample Question



About the optimization of Logistic regression, which statement(s) is true?

- A. The optimization problem is convex.
- B. The optimization problem is non-convex.
- C. If the global minimum exists, then gradient descent is guaranteed to converge to it.
- D. Even if the global minimum exists, gradient descent may not converge to it.

Sample Question



For a soft-margin SVM with the Lagrange multiplier α_i , which statement(s) is incorrect?

- A. The values of α_i for non-support vectors are 0.
- B. SVM optimization is a quadratic optimization problem.
- C. The hyperplane does not change as we move from a linear kernel to higher order polynomial kernels.

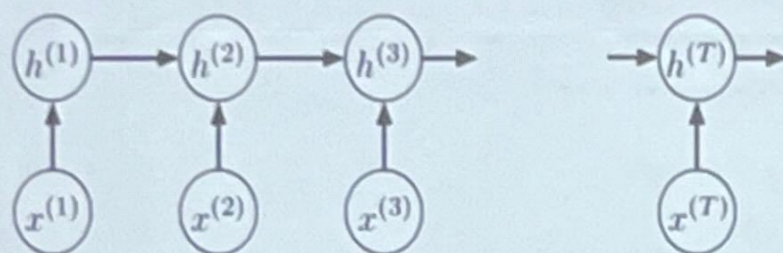
Sequential modeling

- RNNs, LSTMs
 - ▶ Back-propagation through time
 - ▶ Gradient explosion/vanishing
- Transformers
 - ▶ Attention mechanism
 - ▶ Key, Query, Value
- Derive gradient. Decide which gates should be open if we want to copy an input value from t to an output value at $t+m$; Compute attention values for special cases (single layer, same K/Q/V).

Example:

Q4 RNN

2 Points



Q4.1 RNN 1 (Medium)

1 Point

Given a simple univariate RNN with weight w and activation function $\phi(z^t) = \sigma(z^t) - 0.5$ where σ is the sigmoid function and z^t be the input to the activation function at t step. What is the derivative \bar{h}^t for $t < T$?

- ☐ $\bar{h}^t = \bar{h}^{t+1} \sigma'(z^t)$
- ☐ $\bar{h}^t = \bar{h}^{t+1} \sigma'(z^t) w - 0.5$
- ☐ $\bar{h}^t = \bar{h}^{t+1} \sigma'(z^{t+1}) w$
- ☒ $\bar{h}^t = \bar{h}^{t+1} \sigma'(z^t) w$

Solutions: 1. Write down the forward function: $h^{t+1} = \phi(wh^t)$. 2. take the gradient with backprop $\frac{\partial L}{\partial h^t} = \frac{\partial L}{\partial h^{t+1}} \frac{\partial h^{t+1}}{\partial h^t}$.

Example:

Q6 Attention (Medium)

1 Point

Given $V = \begin{bmatrix} 1 & 0 & -1 \\ 2 & -1 & 0 \end{bmatrix}$, $Q = \begin{bmatrix} 1 & 1 & -1 & 1 \\ 0 & -2 & 1 & 1 \end{bmatrix}$, and $K = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix}$, what is the output of the attention layer?

- ☐ $\begin{bmatrix} 1.27 & -0.27 & -0.73 \\ 1.73 & -0.73 & -0.27 \end{bmatrix}$
- ☒ $\begin{bmatrix} 1.5 & -0.5 & -0.5 \\ 1.73 & -0.73 & -0.27 \end{bmatrix}$
- ☐ $\begin{bmatrix} 1.5 & -0.5 & -0.5 \\ 1.88 & -0.88 & -0.12 \end{bmatrix}$
- ☐ $\begin{bmatrix} 1.38 & -0.38 & -0.62 \\ 1.62 & -0.62 & -0.38 \end{bmatrix}$

Solutions: simply apply $A = \text{softmax}(QK^T / \sqrt{D_Q})$



Midterm Practice Problems

March 4, 2024

1. Is it possible to use a linear regression model for binary classification? If so, how do we map the regression output $\mathbf{w}^\top \mathbf{x}$ to the class labels $y \in \{-1, 1\}$?
2. We are given a dataset $\mathcal{D} = \{(-1, -1), (1, 1), (2, 1)\}$ containing three pairs (x, y) , where each $x \in \mathbb{R}$ denotes a real-valued point and $y \in \{-1, +1\}$ is the point's class label.

We want to train the parameters $\mathbf{w} \in \mathbb{R}^2$ (i.e., weight w_1 and bias w_2) of a logistic regression model

$$p(y|x) = \frac{1}{1 + \exp\left(-y\mathbf{w}^\top \begin{bmatrix} x \\ 1 \end{bmatrix}\right)} \quad (1)$$

using maximum likelihood while assuming the samples in the dataset \mathcal{D} to be i.i.d. Instead of maximizing the likelihood we commonly minimize the negative log-likelihood. Specify the objective for the model given in eq. (1). Don't use any regularizer or weight-decay.

3. Consider a model with the following parameterization:

$$p(y^{(i)}|\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x}^{(i)} - b)}, \quad (2)$$

where $\mathbf{w} \in \mathbb{R}^2$ and $b \in \mathbb{R}$.

We want to train the parameters $w \in \mathbb{R}^2$ (i.e., weight w_1 and bias w_2) of a logistic regression model

$$p(y|x) = \frac{1}{1 + \exp\left(-yw^\top \begin{bmatrix} x \\ 1 \end{bmatrix}\right)} \quad (1)$$

using maximum likelihood while assuming the samples in the dataset \mathcal{D} to be i.i.d. Instead of maximizing the likelihood we commonly minimize the negative log-likelihood. Specify the objective for the model given in eq. (1). Don't use any regularizer or weight-decay.

3. Consider a model with the following parameterization:

$$p(y^{(i)}|\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x}^{(i)} - b)}, \quad (2)$$

where $\mathbf{w} \in \mathbb{R}^2$ and $b \in \mathbb{R}$.

What is the highest accuracy for this model on the XOR dataset? **Note:** To compute accuracy, we use a threshold of 0.5, i.e., the final prediction of the model is $\delta[p(y^{(i)}|\mathbf{x}) > 0.5]$, where δ denotes the indicator function.

4. Consider another model with the parametrization shown below:

$$\tilde{y}^{(i)} = \frac{1}{1 + \exp(-a_2^{(i)})} \quad (3)$$

$$a_2^{(i)} = \theta^\top \max(\mathbf{a}_1^{(i)}, 0) + b \quad (4)$$

$$\mathbf{a}_1^{(i)} = \mathbf{W}\mathbf{x}^{(i)} + \mathbf{c} \quad (5)$$

where $\theta \in \mathbb{R}^2$, $b \in \mathbb{R}$, $\mathbf{W} \in \mathbb{R}^{2 \times 2}$ and $\mathbf{c} \in \mathbb{R}^2$.