STA314H1F FINAL PRACTICE QUESTIONS

November 2025

1. Unsupervised Learning

The K-means algorithm aims to find the cluster centroids $\{\mathbf{m}_k\}_{k=1}^K$ and assignments $\{\mathbf{r}^{(n)}\}_{n=1}^N$ to minimize the sum of squared distances of data points $\{\mathbf{x}^{(n)}\}_{n=1}^N$ to their assigned clusters. The objective function is given by:

$$\min_{\{\mathbf{m}_k\},\{\mathbf{r}^{(n)}\}} \hat{\mathcal{R}}(\{\mathbf{m}_k\},\{\mathbf{r}^{(n)}\}) = \min_{\{\mathbf{m}_k\},\{\mathbf{r}^{(n)}\}} \sum_{n=1}^{N}\sum_{k=1}^{K} r_k^{(n)} \|\mathbf{m}_k - \mathbf{x}^{(n)}\|^2$$

where $r_k^{(n)} = \mathbb{I}[\mathbf{x}^{(n)} \text{ is assigned to cluster } k]$

Question 1

Denote $N_k \coloneqq \sum_{n=1}^N r_k^{(n)}$ as the number of points assigned to cluster k. Denote the mean for each cluster k as $\bar{\mathbf{x}}_k = \frac{1}{N_k}\sum_{n=1}^N r_k^{(n)}\mathbf{x}^{(n)}$, and the overall mean of data points as $\bar{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^N \mathbf{x}^{(n)}$. Then, define the following measures:

- Within-cluster scatter: $W(K) = \sum_{k=1}^K \sum_{n=1}^N r_k^{(n)} \|\bar{\mathbf{x}}_k - \mathbf{x}^{(n)}\|^2$
- Between-cluster scatter $B(K) = \sum_{k=1}^K N_k \|\bar{\mathbf{x}}_k - \bar{\mathbf{x}}\|^2$
- Total point scatter: $T = \sum_{n=1}^N \|\bar{\mathbf{x}} - \mathbf{x}^{(n)}\|^2$

a) Prove that $T = \frac{1}{2N}\sum_{i=1}^N \sum_{j=1}^N \|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|^2$

b) Prove that $W(K) = \frac{1}{2}\sum_{k=1}^K \frac{1}{N_k}\sum_{i=1}^N\sum_{j=1}^N r_k^{(i)} r_k^{(j)} \|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|^2$. You may use the result of a).

c) Show that $W(K) + B(K) = T$

Question 2 (From lecture)

Show that as $\beta \to \infty$, soft K-means becomes K-means. (Hint: analyze the case when k is the index of the closest centroid and vice versa.)

Question 3

When using K-means for vector quantization on images, we construct the dataset of pixels denoted by $X \in \mathbb{R}^{N\times 3}$, where the rows represent the N pixels and columns represent the RGB intensities. Suppose we run K-means (with K clusters) on X and store the cluster centroids $\mathbf{m}_k$ in the rows of matrix M, and store the cluster assignments $\mathbf{r}_k^{(n)}$ in the rows of matrix Z.

a) What would be the dimensions of M and Z?

b) Using M and Z, how can we construct a low-rank approximation of X, call it $\hat{X}$?

c) What can you say about the compressed (low-rank) image?

2. PCA

(a) True or False: PCA is a supervised learning method.

(b) Show that the reconstruction of the data, $\tilde{\boldsymbol{x}}$, has the same mean as the original data $\boldsymbol{x}$

(c) Why do we typically scale and center the data prior to performing PCA? What are potential downsides if this step is skipped?

(d) What metric do we wish to minimize when choosing a subspace? Is this equivalent to another metric?

(e) Show that if A is an orthogonal matrix, then it's determinant is either 1 or -1.

(f) What do the eigenvalues of a covariance matrix represent? What does the largest eigenvalue of a covariance matrix correspond to?

3. Probabilistic modelling

You flip a coin twice and observe 2 heads. Let θ denote the probability of a given flip showing heads.

(a) Calculate $\hat{\theta}_{\text{MLE}}$, the MLE for θ. How does this value go against your intuition?

(b) Assume a prior $\theta \sim \text{Beta}(2, 2)$. Find the posterior distribution $p(\theta|\mathcal{D})$. (Hint: The density of the Beta(α, β) distribution is $\propto x^{\alpha-1}(1-x)^{\beta-1}$).

(c) Calculate $\hat{\theta}_{\text{MAP}}$, the MAP estimator under the prior in part (b).

(d) Calculate the posterior mean $\mathbb{E}[\theta|\mathcal{D}]$.

(e) In a sentence or two each, explain:

  (i) Why the MLE can catastrophically overfit under data sparsity.

  (ii) How a Bayesian approach and MAP can smooth predictions and where may they still fail.

(f) Repeat questions (b) and (c) under a Beta$(1, 1)$ prior. What do you observe?

4. Linear Regression

(a) True or False: The linear regression prediction $y = w^\top x + b$ is considered linear in x but it is nonlinear in w and b.

(b) Write the average squared loss $\hat{R}$ in vectorized form using the design matrix X, weight vector $\mathbf{w}$, bias b, and target vector t.

(c) State the dimensions of the augmented design matrix $\tilde{X}$ (augmented means we add a dummy feature always equal to 1) and the augmented weight vector $\tilde{w}$ if the original dataset has N examples and D features.

(d) What does the residual $(y^{(i)} - t^{(i)})$ represent for a single data point, and what is the primary objective of the least squares method with respect to the residuals across the entire dataset?

(e) In the linear model $y = w^\top x + b$, what do we call a specific choice of the parameters (w, b)?

5. Regularized Regression

(a) Explain the problem that L2-regularization (Ridge Regression) addresses regarding models that overfit, specifically in relation to the magnitudes of the model coefficients.

(b) Write the closed-form solution for the Ridge Regression weights $w^{\lambda}_{\text{Ridge}}$, using the design matrix X (unaugmented) and the hyperparameter λ.

(c) What is the primary function of the hyperparameter λ in the regularized loss $\hat{R}_{\text{reg}}(w) = \hat{R}(w) + \lambda\phi(w)$, and how is its optimal value typically determined in practice?

(d) The Gradient Descent update for the L2-regularized loss results in the final term $\mathbf{w} \leftarrow (1 - \alpha\lambda)\mathbf{w} - \alpha\frac{\partial \hat{R}}{\partial \mathbf{w}}$. What is the common name for the effect of the multiplicative term $(1 - \alpha\lambda)$, and what characteristic of the weights does it enforce?

(e) How can a linear regression model be extended to perform non-linear fitting, such as polynomial regression, without altering the core linear regression algorithm itself?

6. Bayesian Linear Regression

(a) True or False: Using a Bayesian implementation of linear regression can help prevent overfitting problems that easily arise for maximum likelihood.

(b) Consider a Bayesian linear regression setting with the vector of targets $\mathbf{t}$, the matrix of inputs $\mathbf{X}$, design matrix $\mathbf{\Psi} = \mathbf{\Psi}(\mathbf{X})$, and σ^2 known. The likelihood function for the target vector is

$$p(\mathbf{t} \mid \mathbf{X}, \mathbf{w}) \sim \mathcal{N}(\mathbf{\Psi}(\mathbf{X})\mathbf{w},\, \sigma^2\mathbf{I}),$$

and the weight vector prior is

$$p(\mathbf{w}) \sim \mathcal{N}(\boldsymbol{\mu_0},\, \mathbf{\Sigma_0}).$$

(i) What is the maximum posterior (MAP) weight vector?
Note: You may use standard results from Bayesian linear regression without deriving them, but you should clearly explain the reasoning you use and justify the results you obtain.

(ii) Now suppose new inputs $\mathbf{X_1}$ and targets $\mathbf{t_1}$ become available. What is the new MAP weight vector? How does it defer from the one found in (a)?

(iii) One last vector of inputs, $\mathbf{x_{new}}$, becomes available and we want to predict t_{new}, the target associated with it. Using the samples from (i) and (ii), give an expression for the prediction t_{new}.

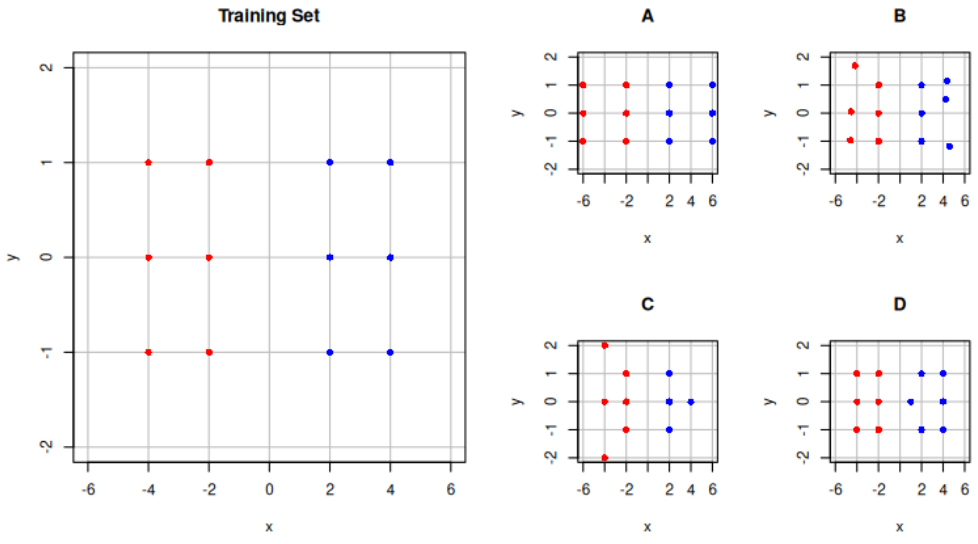
7. Linear Classification

(a) True or False the following dataset is linearly separable.

x_1	x_2	x_3	x_4	t
2	5	8	2	1
2	15	7	-1	1
4	25	8	-7	1
-2	10	4	-5	0
-3	11	9	-3	0

(b) An SVM is fit to the training data $\{(\boldsymbol{x}^{(i)}, t^{(i)})\}_{i=1}^n$, where $\boldsymbol{x}^{(i)} = (x_1^{(i)}, x_2^{(i)})$ and $t^{(i)} \in \{0, 1\}$. A new dataset is constructed by scaling the first feature, $\tilde{\boldsymbol{x}}^{(i)} = (cx_1^{(i)}, x_2^{(i)})$ and a different SVM is fit on the new dataset $\{(\tilde{\boldsymbol{x}}^{(i)}, t^{(i)})\}_{i=1}^n$.
True or False, if $\boldsymbol{x}^{(i)}$ is a support vector of the first SVM, then $\tilde{\boldsymbol{x}}^{(i)}$ is always a support vector of the second SVM.

(c) Suppose we have an SVM fit using the training set shown below on the left. Which of the training sets shown below on the right would produce the same SVM.



(d) Show that the decision boundary for a logistic regression model is linear.

8. Matrix Factorization

(a) Let $\boldsymbol{A} \in \mathbb{R}^{d \times d}$. Prove that

$$||\boldsymbol{A}||_F^2 = \text{tr}(\boldsymbol{A}^\top \boldsymbol{A}).$$

It is known that

$$\text{tr}(\boldsymbol{M}) = \lambda_i(\boldsymbol{M})$$

where $\lambda_1(\boldsymbol{M}), \ldots, \lambda_d(\boldsymbol{M})$ denote the eigenvalues of the matrix $\boldsymbol{M}$ in non-increasing order. Based on this, we can conclude (you don't need to prove this) that

$$||\boldsymbol{A}||_F^2 = \sum_{i=1}^{d} \lambda_i(\boldsymbol{A}^\top \boldsymbol{A}).$$

(b) Assume $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is symmetric and positive semi-definite (PSD). In particular, it has eigenvalues $\lambda_1, \ldots, \lambda_d \geq 0$ with associated (normalized) eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d \in \mathbb{R}^d$. It can be compactly decomposed as $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ such that $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_d)$ and $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d] \in \mathbb{R}^{d \times d}$.

Now, define $\boldsymbol{Z} = \boldsymbol{U}_k \boldsymbol{\Lambda}_k^{1/2}$, where $\boldsymbol{U}_k = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k] \in \mathbb{R}^{d \times k}$ and $\Lambda_k = \text{diag}(\lambda_1, \ldots, \lambda_k) \in \mathbb{R}^{k \times k}$. Derive the reconstruction error

$$||\boldsymbol{A} - \boldsymbol{Z}\boldsymbol{Z}^\top||_F^2$$

for the rank-k approximation of $\boldsymbol{A}$ by $\boldsymbol{Z}\boldsymbol{Z}^\top$.

9. Bias, Variance, and Ensembles

We have N scalar-valued observations $\mathcal{D}^{train} = \{x^{(i)}\}_{i=1}^{N}$ sampled independently from a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$. We want to find an estimator, $\hat{\mu}$, to estimate the true mean μ. In this problem, we will go over several estimators, and focus on the **Mean Squared Error (MSE)** of these estimators: $\text{MSE}(\hat{\mu}) = \mathbb{E}[(\hat{\mu} - \mu)^2]$.

- **Question 0: Decomposition of MSE**

  Show that:

  $$\text{MSE}(\hat{\mu}) = \text{Bias}^2 + \text{Variance}$$

  where $\text{Bias} = \mathbb{E}[\hat{\mu}] - \mu$ and $\text{Variance} = \text{Var}[\hat{\mu}]$

- **Question 1: Simple Estimators**

  a) First, consider a simple estimator that only uses the first data point: $\hat{\mu}_1 = x^{(1)}$. Calculate its Bias^2, Variance, and $\text{MSE}(\hat{\mu}_1)$.

  b) Now, recall the standard sample mean estimator: $\hat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}$. State its Bias^2, Variance, and $\text{MSE}(\hat{\mu}_N)$. Briefly comment on why this MSE is different from that of $\hat{\mu}_1$.

- **Question 2: Regularized Estimator**

  Now, consider a new estimator $\hat{\mu}_\lambda$. Let $\hat{\mu}_\lambda = \frac{N}{N+\lambda}\hat{\mu}_N = \left(\frac{1}{N+\lambda}\right)\sum_{i=1}^{N} x^{(i)}$, where $\lambda \geq 0$ is a fixed hyperparameter.

  a) Calculate the Bias^2 of this estimator, $\text{Bias}^2(\hat{\mu}_\lambda)$. Whether it's unbiased estimator?

  b) Calculate the Variance of this estimator, $\text{Var}(\hat{\mu}_\lambda)$.

  c) Using your results, find the optimal λ^* that **minimizes** the $\text{MSE}(\hat{\mu}_\lambda)$. (You may assume $\mu \neq 0$ and $\sigma^2 > 0$).

- **Question 3: Bagged Estimator**

  Bagging (Bootstrap Aggregation) is an ensemble technique used to reduce variance. Let's analyze its mechanics for a very small dataset.

  Assume our entire training set $\mathcal{D}^{train} = \{x^{(1)}, x^{(2)}\}$. Our base estimator is the sample mean of a dataset $\mathcal{D}$: $\hat{\mu}(\mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{j\in\mathcal{D}} x^{(j)}$. We will create bootstrap samples $\mathcal{D}_{boot}$ by sampling $N = 2$ points with replacement from $\mathcal{D}^{train}$.

  a) List all 4 possible bootstrap samples $\mathcal{D}_{boot,k}$.

b) For each of the 4 samples, compute its estimate $\hat{\mu}_k = \hat{\mu}(\mathcal{D}_{boot,k})$ in terms of $x^{(1)}$ and $x^{(2)}$.

c) The bagged prediction $\hat{\mu}_{bag}$ is the average of the predictions from all 4 bootstrap samples. Compute $\hat{\mu}_{bag}$ in terms of $x^{(1)}$ and $x^{(2)}$. What do you observe about $\hat{\mu}_{bag}$ compared to $\hat{\mu}_N$ (for $N = 2$) from Question 1b?