STAT 154/254 Midterm

Raymond Tsao

TOTAL POINTS

80 / 101

QUESTION 1

Multiple choice 52 pts

1.1 1 **4 / 4**

✓ **- 0 pts** *Correct*

- 4 pts Wrong

1.2 2 **4 / 4**

✓ **- 0 pts** *Correct*

- 1 pts Logistic and Linear are included in the answer, though the link function itself is "most" associated with GLM itself

- 4 pts Wrong

1.3 3 **4 / 4**

✓ **- 0 pts** *Correct*

- 4 pts Wrong

1.4 4 **0 / 4**

- 0 pts Correct

✓ **- 4 pts** *Wrong*

1.5 5 **4 / 4**

✓ **- 0 pts** *Correct*

- 4 pts Wrong

1.6 6 **4 / 4**

✓ **- 0 pts** *Correct*

- 4 pts Wrong

1.7 7 **4 / 4**

✓ **- 0 pts** *Correct*

- 4 pts Wrong

1.8 8 **0 / 4**

- 0 pts Correct

✓ **- 4 pts** *Wrong*

1.9 9 **4 / 4**

✓ **- 0 pts** *Correct*

- 4 pts Wrong

1.10 10 **4 / 4**

✓ **- 0 pts** *Correct*

- 4 pts Wrong

1.11 11 **0 / 4**

- 0 pts Correct

✓ **- 4 pts** *Wrong*

1.12 12 **4 / 4**

✓ **- 0 pts** *Correct*

- 4 pts Wrong

1.13 13 **4 / 4**

✓ **- 0 pts** *Correct*

- 4 pts Wrong

QUESTION 2

2 Logistic Regression 10 / 10

✓ **- 0 pts** *Correct*

- 10 pts Blank/Wrong

- 7 pts Only a) is correct

- 8 pts Partially correct arguments for both, but questions are not properly answered/explained

- 2 pts Not enough details/some problems

QUESTION 3

3 Support Vector Machines 15 / 15

+ 0 pts Blank / no progress.

✓ **+ 5 pts** *Part (a) correct*

✓ **+ 5 pts** *Part (b) correct*

✓ **+ 5 pts** *Part (c) correct*

QUESTION 4

4 Linear Discriminant Analysis 10 / 10

✓ **- 0 pts** *Correct*

- 10 pts Blank / no progress.

- 1 pts Small calculation mistakes

- 6 pts Derivation missing

- 4 pts (b) not attempted

QUESTION 5

5 MLE of OLS 5 / 14

- 0 pts Correct

- 14 pts Blank / no progress.

- 4 pts Significant progress.

- 8 pts Some progress

✓ **- 10 pts** *Some correct conclusions.*

+ 1 *Point adjustment*

Stat154/254 Modern Statistical Prediction and Machine Learning

Midterm Exam

Thursday, October 5, 2023

Instructor: Nikita Zhivotovskiy **GSI**: Mriganka Basu Roy Chowdhury

Maximum Points: 87 (for STAT 154), 101 (for STAT 254) **Duration**: 80 minutes.

Write your name: Raymond Tsao

SID Number: 3037860126

Please select the course you are enrolled in

■ STAT 154. ☐ STAT 254.

Exam Information and Instructions:

- This is a closed book, closed notes exam. The use of electronic devices is not permitted.
- Write any work you want graded on the front of each page. The reverse side of the page can be used as scratch paper. Additionally, write your SID number in the top right corner on every **extra** page you might need.
- Provide reasoning for all **Written Questions**.
- For multiple answer questions, fill in the answer for **ALL** correct choices. At least one of the answers is always correct. **NO** partial credit for the multiple answer questions.
- Please write your answers as clearly and legibly as possible.

I certify that all materials in the enclosed exam are my own original work and I followed Berkeley Honor Code.

Sign your name: Raymond Tsao

Good Luck!

1. Multiple Answer Questions (52 points)

For multiple answer questions, fill in the answer for **ALL** correct choices. **NO** partial credit on multiple answer questions. Multiple choice questions are all worth 4 **points**.

1. In the context of linear regression, which abbreviations are associated with variable selection?

- ☐ RSS: Regularized Stratification Strategy.
- ■ LASSO: Least Absolute Shrinkage and Selection Operator.
- ☐ TSS: Targeted Shrinkage Selector.
- ☐ RSE: Regression Selection Estimator.

2. The term "link function" is most associated with:

- ☐ SVM.
- ☐ Logistic Regression.
- ■ Generalized Linear Model.
- ☐ Linear Regression.

3. In a two-class problem where each class follows a Gaussian distribution with different means but the same covariance, how does the Naive Bayes classifier (diagonal) compare to Linear Discriminant Analysis (LDA)?

- ■ For non-diagonal covariance, Naive Bayes and LDA yield different decision boundaries due to the feature independence assumption in Naive Bayes.
- ■ For diagonal covariance, both classifiers provide the same decision boundary.
- ☐ LDA always assumes feature independence.
- ☐ Both classifiers always produce identical results, regardless of the covariance structure.

4. Which of the following cross-validation versions may not be suitable for large datasets with hundreds of thousands of samples?

- ☐ k-fold cross-validation when k is close to the sample size.
- ☐ Holdout method using a random 50%-50% train/test split.
- ■ Leave-one-out cross-validation.
- ☐ Holdout method using a random 80%-20% train/test split.

5. You are employing a regularized least squares regression model and progressively increasing the regularization parameter. Which of the following behaviors is typically observed as the regularization increases?

- ☐ The model complexity increases, leading to a more flexible fit.
- ☐ Test error drops consistently as regularization increases.
- ■ Bias typically increases while variance decreases.
- ■ The weight coefficients of the model tend to zero.

6. In the context of linear regression, when testing for the significance of a predictor's coefficient using the t-statistic, which of the following statements is correct?

- ■ A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis that the predictor's coefficient is zero.
- ☐ A large p-value suggests that the relationship between the predictor and the response variable is strong.
- ■ The t-statistic represents the number of standard errors by which the coefficient estimate differs from zero.
- ■ The t-statistic is computed as the ratio of the coefficient estimate to its standard error. $\frac{\hat{\beta}_0}{SE(\hat{\beta}_0)}$

$y = 50 + 10X_1 + 20X_2 - 5X_1X_2$

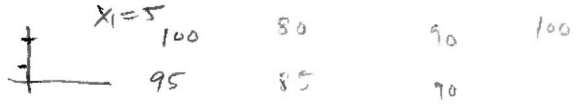
7. Given a dataset with predictors X_1 (GPA), X_2 (Level, 1 denotes College and 0 denotes High School), and X_3 (Interaction between GPA and Level, that is, the product of two predictors), the response is the starting salary after graduation (in thousands of dollars). Using least squares, the coefficients obtained are: $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 10$, $\hat{\beta}_2 = 20$, and $\hat{\beta}_3 = -5$.

Which of the following statements is accurate based on the coefficients?

- ☐ High school graduates always earn more than college graduates, irrespective of GPA.
- ■ High school graduates earn more than college graduates, but only if GPA is sufficiently high.
- ☐ College graduates always earn more than high school graduates, irrespective of GPA.
- ☐ College graduates earn more than high school graduates, but only if GPA is sufficiently high.

no college $y = 50 + 10X_1$

w/ college $y = 50 + 10X_1 + 20 - 5X_1 = 70 + 5X_1$



8. In the context of linear regression, which of the following statements are true regarding leverage scores and outliers?

- ☐ Outliers always distort the regression line significantly.
- ☐ An observation can be an outlier without having a high leverage score.
- ■ Leverage scores specifically quantify how much an observation's features differ from the mean of all features.
- ■ Observations with high leverage scores can heavily influence the slope of the regression line.

9. In polynomial regression, which of the following assumptions most directly affects the trade-off between underfitting and overfitting?

- [x] Degree of the polynomial.
- [ ] Doubling the size of the test sample.
- [ ] Choice of a regularization term (e.g., $\|\cdot\|_1$ or $\|\cdot\|_2$).
- [ ] Choice of optimization method (e.g., matrix inversion vs. Newton's method).

10. Consider a binary classification problem with classes labeled as $\{1, -1\}$ and a feature vector $x \in \mathbb{R}^p$. In the Perceptron algorithm, with the initial intercept set to zero and kept constant, which formula correctly describes the weight update rule for a misclassified instance?

- [x] $w_{\text{new}} = w_{\text{old}} + y \cdot x$
- [ ] $w_{\text{new}} = w_{\text{old}} - y \cdot x$
- [ ] $w_{\text{new}} = w_{\text{old}} + \frac{y \cdot \langle w_{\text{old}}, x \rangle}{\|w_{\text{old}}\|_2}$
- [ ] $w_{\text{new}} = w_{\text{old}} - \frac{y \cdot \langle w_{\text{old}}, x \rangle}{\|w_{\text{old}}\|_2}$

$y\langle w_{new}, x\rangle = y\langle w_{old}, x\rangle + 1\|x\|^2$

11. Which of the following statements compare hard margin SVMs with Perceptrons?

- [x] Both find a linear decision boundary to separate classes.
- [ ] SVMs maximize the margin between classes unlike Perceptrons.
- [ ] Both use an iterative update rule.
- [ ] Perceptrons can work with non-linearly separable data.

12. Which of the following statements best describes the advantage of using Lasso rather than ridge regression?

- [ ] Lasso performs continuous shrinkage of coefficients like ridge regression.
- [ ] Lasso results in lower prediction error compared to ridge regression.
- [ ] Lasso completely eliminates the least important variables from the model.
- [x] Lasso identifies relevant variables by shrinking some coefficients to zero.

1 $\quad \frac{r^2}{\gamma^2} \quad \frac{r^2}{\gamma^2}$ Stops in at most $\frac{r^2}{\gamma^2}$ iterations $\Rightarrow \frac{r^2}{\gamma^2} - 1$ errors

13. Which of the following can be inferred from Novikoff's theorem on the number of mistakes made by the perceptron algorithm, given that the data points are linearly separable and have the Euclidean norm at most 1?

$\frac{1}{\gamma^2}$ errors

- [x] For linearly separable data with a non-zero margin, the perceptron algorithm guarantees convergence in a finite number of steps.
- [ ] Larger margin usually leads to larger number of mistakes.
- [ ] The number of mistakes equals the number of training examples.
- [x] The number of mistakes is finite for linearly separable data with non-zero margin.

Written Questions.

2. **Logistic Regression** (3 + 7 = 10 points).

Consider a binary classification problem with classes labeled as $\{0, 1\}$, where you have a real-valued feature x. You aim to construct a classifier using a simple logistic regression model.

(a) Describe the procedure to construct a binary classifier based on the estimated conditional probability $\Pr(Y = 1|X = x)$.

(b) Formally demonstrate that the logistic regression model generates a linear decision boundary.

a) Model $\mathbb{P}\{Y=1 \mid X=x\} = \frac{1}{1+\exp(-\beta^T x)}$ if $x \in \mathbb{R}^p$, then $\beta \in \mathbb{R}^p$

one can add a 1 to all feature vectors at the beginning to model the bias term

We then estimate $\hat{\beta}$ by Maximizing the likelihood, this can be done w/ gradient descent or Newton's method

After $\hat{\beta}$ is estimated, we compare $\mathbb{P}\{Y=1 \mid X=x\} = \frac{1}{1+\exp(-\beta^T x)}$ and $\mathbb{P}\{Y=0 \mid X=x\} = \frac{\exp(-\beta^T x)}{1+\exp(-\beta^T x)}$ for each test input and predict the class that gives a higher probability

b) The decision boundary is $\mathbb{P}\{Y=1 \mid X=x\} = \mathbb{P}\{Y=0 \mid X=x\}$

or $\frac{\exp(-\beta^T x)}{1+\exp(-\beta^T x)} = \frac{1}{1+\exp(-\beta^T x)} \iff \exp(-\beta^T x) = 1$

$\iff -\beta^T x = 0 \iff \beta^T x = 0$

which is a linear function.

$\prod_{i:y_i=1} \left(\frac{1}{1+\exp(-\beta^T x_i)}\right) \prod_{i:y_i=0} \left(\frac{\exp(-\beta^T x_i)}{1+\exp(-\beta^T x_i)}\right)$ or alternatively,

log-likelihood

$$\sum_{i:y_i=1} -\log(1+\exp(\beta^T x_i)) + \sum_{i:y_i=0} -\beta^T x_i - \log(1+\exp(\beta^T x_i))$$

3. **Support Vector Machines (SVM) - Hard Margin** (5 + 5 + 5 = 15 points).

Consider a binary classification problem in a two-dimensional space with linearly separable classes labeled as $\{-1, 1\}$. You intend to classify these points using a hard margin SVM. Let the training sample be represented as pairs (x_i, y_i) for $i = 1, \ldots, N$. The decision boundary (or halfspace) is characterized by the slope vector β and the intercept β_0. The **Karush-Kuhn-Tucker (KKT) Conditions** for the SVM are as follows:

$$\beta = \sum_{i=1}^{N} \alpha_i y_i x_i, \quad \sum_{i=1}^{N} \alpha_i y_i = 0, \quad \alpha_i \geq 0,$$
$$\alpha_i \left[y_i (x_i^T \beta + \beta_0) - 1 \right] = 0 \quad \text{for all } i = 1, \ldots, N.$$

(a) If you have three data points x_1, x_2, and x_3 such that their corresponding $\alpha_1 > 0$, $\alpha_2 = 0$, and $\alpha_3 > 0$, which of these data points lie on the boundary of the slab (margin)? Justify your answer using the KKT conditions.

(b) Using the KKT conditions and the computed dual variables $\alpha_1, \ldots, \alpha_N$, describe the method to compute the slope vector β for the decision boundary of the SVM. How would you compute the intercept β_0 using the SVM model? Provide a step-by-step process. Your answers may depend on the computed values $\alpha_1, \ldots, \alpha_N$.

(c) Suppose you apply the trained SVM classifier to a test sample. Define a binary loss such that the loss is equal to 0 if the point is classified correctly and 1 otherwise. Formally define the classification error for the test sample and describe a method to compute the classification error rate using the SVM model.

a) Recall that at point x_i, the margin γ is given by $y_i(x_i^T\beta+\beta_0) \geq 1$ (Assuming the weights β_i and β_0 are normalization). Equility holds if x_i is on the line of margin.

Since $\alpha_1, \alpha_3 > 0$. It follows that $y_i(x_i^T\beta+\beta_0)=1$ for $i=1,3$, this shows that x_1, x_3 lie on the boundary

b) since x_i, y_i are known for $i=1,2,\ldots N$, and $\alpha_1 \ldots \alpha_N$ are computed, β can be computed by $\beta = \sum_{i=1}^{N} \alpha_i y_i x_i$ ← all known variables

once β is estimated, we can determine β by using the condition

$$\alpha_i [y_i(x_i^T\beta+\beta_0)-1]=0$$

for those whose $\alpha_i > 0$, solve β_0 by solving

$$y_i x_i^T \beta + y_i \beta_0 = 1$$

This gives us β_0 since x_i, y_i are known

$\Sigma = \Sigma^T$
$\Sigma^{-1} = \Sigma^{T^{-1}}$

Use the front of this page for solutions that don't fit on the designated page.

c) Binary loss

$\ell(y_{pred}, y_{true}) \overset{def}{=} \mathbb{1}\{y_{pred} = y_{true}\}$ $\quad y_{pred} = f(X)$ ← model

we want to estimate $\mathbb{E}[\ell(f(X), Y)]$ where expectation is taken under population distribution. To do so, given test data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, define test error as

$$\frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i) \overset{\text{estimate}}{\rightsquigarrow} \mathbb{E}[\ell(f(X), Y)]$$

$$\| $$

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{f(x_i) = y_i\}$$

where f is the classification model.

To compute the classication error rate, for each input x_i, we classify x_i based on which side it lies on the decision boundary. (since SVM partitions the feature space using the decision boundary solved in part (b))

This gives us y prediction, we can compute classification error rate w/ the formula

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{y_{pred,i} = y_{true,i}\}$$

5. **Linear Discriminant Analysis (LDA)** (6 + 4 = 10 points).

Consider a binary classification problem in a two-dimensional space with classes labeled as $\{-1, 1\}$. You wish to utilize LDA for classification and thus assume that classes have different expectations μ_1, μ_2 but the same covariance matrix Σ. Recall the density function for the multivariate Gaussian distribution:

$$f(x) = \frac{1}{(2\pi)^{\frac{p}{2}}(\det(\Sigma))^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right),$$

where p is the number of features, Σ is the covariance matrix, and μ is the mean vector.

(a) Given the estimated parameters for both classes, derive the LDA decision rule to classify a new point x_{new}, provided that prior probabilities of classes are p_- and p_+ respecitively.

(b) Consider two classes with the following estimated parameters:

$$\widehat{\mu}_1 = \begin{bmatrix}0\\3\end{bmatrix}, \quad \widehat{\mu}_2 = \begin{bmatrix}3\\6\end{bmatrix}, \quad \widehat{\Sigma} = \begin{bmatrix}2 & 1\\1 & 2\end{bmatrix}, \quad \widehat{\Sigma}^{-1} = \frac{1}{3}\begin{bmatrix}2 & -1\\-1 & 2\end{bmatrix}.$$

Assume that prior probabilities of both classes are equal. Compute the LDA projection of the point $x_{\text{new}} = \begin{bmatrix}1\\4\end{bmatrix}$ and classify it.

$\frac{e^a}{e^b} = e^{a-b}$

a) $\mathbb{P}\{Y=1 \mid X=x\} = \dfrac{\mathbb{P}\{X=x \mid Y=1\}\,\mathbb{P}\{Y=1\}}{\mathbb{P}\{X=x\}}$

$\mathbb{P}\{Y=-1 \mid X=x\} = \dfrac{\mathbb{P}\{X=x \mid Y=-1\}\,\mathbb{P}\{Y=-1\}}{\mathbb{P}\{X=x\}}$

$\mathbb{P}\{Y=1 \mid X=x\} > \mathbb{P}\{Y=-1 \mid X=x\}$ if $\mathbb{P}\{X=x \mid Y=1\}\,p_+ > \mathbb{P}\{X=x \mid Y=-1\}\,p_-$

$\Leftrightarrow \dfrac{\exp(-\frac{1}{2}(x-\mu_1)^\top\Sigma^{-1}(x-\mu_1))}{\exp(-\frac{1}{2}(x-\mu_2)^\top\Sigma^{-1}(x-\mu_2))} > \dfrac{p_-}{p_+}$

$\Leftrightarrow \exp(-\frac{1}{2}(x-\mu_1)^\top\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)^\top\Sigma^{-1}(x-\mu_2)) > \dfrac{p_-}{p_+}$

$\Leftrightarrow -(x-\mu_1)^\top\Sigma^{-1}(x-\mu_1) + (x-\mu_2)^\top\Sigma^{-1}(x-\mu_2) > 2\log\dfrac{p_-}{p_+}$

$\Leftrightarrow x^\top\Sigma^{-1}x - \mu_2^\top\Sigma^{-1}x - x^\top\Sigma^{-1}\mu_2 + \mu_2^\top\Sigma^{-1}\mu_2 -$
$(x^\top\Sigma x - \mu_1^\top\Sigma^{-1}x - x^\top\Sigma^{-1}\mu_1 + \mu_1^\top\Sigma^{-1}\mu_1) > 2\log\dfrac{p_-}{p_+}$

$\Leftrightarrow \mu_1^\top\Sigma^{-1}x - \mu_2^\top\Sigma^{-1}x + x^\top\Sigma^{-1}\mu_1 - x^\top\Sigma^{-1}\mu_2 + \mu_2^\top\Sigma^{-1}\mu_2 - \mu_1^\top\Sigma^{-1}\mu_1 > 2\log\dfrac{p_-}{p_+}$

let $\boxed{l(x) = \mu_1^\top\Sigma^{-1}x - \mu_2^\top\Sigma^{-1}x + x^\top\Sigma^{-1}\mu_1 - x^\top\Sigma^{-1}\mu_2 + \mu_2^\top\Sigma^{-1}\mu_2 - \mu_1^\top\Sigma^{-1}\mu_1 - 2\log\frac{p_-}{p_+}}$

classify label of x as $+1$ if $l(x) > 0$, else classify as -1

Note that is a linear function of x, can be further simplified (see next)

Use the front of this page for solutions that don't fit on the designated page.

b) under LDA, we have normalization constant given by $\frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{\det\Sigma}}$

$$P\{X=x \mid Y=1\} = \frac{P_+}{P\{X=x\}} \cdot C \cdot \exp\left(-\frac{1}{2}\begin{bmatrix}1 & 1\end{bmatrix}\cdot\frac{1}{3}\begin{bmatrix}2 & -1\\ -1 & 2\end{bmatrix}\begin{bmatrix}1\\ 1\end{bmatrix}\right)$$

since $x-\mu_1 = \begin{bmatrix}1\\ 1\end{bmatrix}$ $\quad \xi$

$$= \xi \exp\left(-\tfrac{1}{3}\right)$$

$$P\{X=x \mid Y=-1\} = \xi \exp\left(-\frac{1}{2}\begin{bmatrix}-2 & -2\end{bmatrix}\frac{1}{3}\begin{bmatrix}2 & -1\\ -1 & 2\end{bmatrix}\begin{bmatrix}-2\\ -2\end{bmatrix}\right)$$

since $x-\mu_2 = \begin{bmatrix}-2\\ -2\end{bmatrix}$

$$= \xi \exp\left(-\tfrac{4}{3}\right)$$

since $P\{X=x \mid Y=1\} > P\{X=x \mid Y=-1\}$

we classify x as 1

$$\ell(x) = (\mu_1-\mu_2)^T\Sigma^{-1}x + x^T\Sigma^{-1}(\mu_1-\mu_2) + \mu_2^T\Sigma^{-1}\mu_2 - \mu_1^T\Sigma^{-1}\mu_1 - 2\log\frac{P_-}{P_+}$$

$$= (\mu_1-\mu_2)^T\Sigma^{-1}x + (\mu_1-\mu_2)^T(\Sigma^{-1})^T x + \mu_2^T\Sigma^{-1}\mu_2 - \mu_1^T\Sigma^{-1}\mu_1 - 2\log\frac{P_-}{P_+}$$

$$= 2(\mu_1-\mu_2)\Sigma^{-1}x + \mu_2^T\Sigma^{-1}\mu_2 - \mu_1^T\Sigma^{-1}\mu_1 - 2\log\frac{P_-}{P_+}$$

6. **MSE of OLS** (14 points) (*This problem is **only for STAT 254** students and does not count for the Max score for STAT 154, but can be solved for extra points.*). Consider the Gaussian linear model:

$$y_i = \langle x_i, \beta \rangle + \varepsilon_i \quad \text{for} \quad i = 1, \ldots, N,$$

where ε_i are independent $\mathcal{N}(0, \sigma^2)$ random variables and $\beta, x_1, \ldots, x_N \in \mathbb{R}^p$. Let $\widehat{\beta}_N$ denote the least squares solution. Show the following bound on the expected MSE:

$$A = \frac{1}{N}\mathbb{E}\sum_{i=1}^{N}\left(\langle x_i, \widehat{\beta}_N \rangle - \langle x_i, \beta \rangle\right)^2 \leq \frac{\sigma^2 p}{N},$$

where the expectation is taken with respect to $\varepsilon_1, \ldots, \varepsilon_N$.

Let $\hat{y}_i$ be the estimated solution $\Rightarrow \hat{y}_i = \langle x_i, \widehat{\beta}_N \rangle$, then

$$A = \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^{N}(\hat{y}_i - y_i - \varepsilon_i)^2\right] = \frac{1}{N}\left[\mathbb{E}\left[\sum_{i=1}^{N}(\hat{y}_i - y_i)^2\right] - 2\mathbb{E}\left[\sum_{i=1}^{n}\varepsilon_i(\hat{y}_i - y_i)\right] + \mathbb{E}\left[\sum_{i=1}^{n}\varepsilon_i^2\right]\right]$$

$$= \frac{1}{N}\left[\sum_{i=1}^{n}\mathbb{E}\left[(\hat{y}_i - y_i)^2\right] - 2\sum_{i=1}^{n}\mathbb{E}[\varepsilon_i \hat{y}_i] - 2\mathbb{E}[\varepsilon_i y_i] + \sigma^2 \cdot N\right]$$

$$= \frac{1}{N}\left[\sum_{i=1}^{N}\mathbb{E}\left[(\hat{y}_i - y_i)^2\right] - 2N\sigma^2 + \sigma^2 N\right]$$