STAT 102B: Sample Exam I Questions

George Michailidis

1 Problems that require calculations

Problem 1:

Consider the function

$$f(x) = (x+2)^2.$$

Use Newton's algorithm to perform **one iteration** starting from $x_0 = 5$.

Answer:

Step 1: Compute the first and second derivatives of $f(x)$:

$$f'(x) = \frac{d}{dx}(x+2)^2 = 2(x+2),$$

$$f''(x) = \frac{d}{dx}\left[2(x+2)\right] = 2.$$

Step 2: Apply Newton's update rule:

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}.$$

Step 3: Plug in $x_0 = 5$:

$$f'(5) = 2(5+2) = 14, \quad f''(5) = 2,$$

$$x_1 = 5 - \frac{14}{2} = 5 - 7 = -2.$$

Problem 2:

Consider the lasso regression problem.

Write pseudo-code that implements the proximal gradient algorithm with a fixed step size η.

Be as detailed as possible.

Answer:

Algorithm 1 Proximal Gradient Algorithm for Lasso (Fixed Step Size + Stopping Criterion)

Require: Design matrix $X \in \mathbb{R}^{n \times p}$, response vector $y \in \mathbb{R}^n$, regularization parameter $\lambda > 0$, step size $\eta > 0$, tolerance **tol**> 0, maximum iterations K

Initialize $\beta_0 \in \mathbb{R}^p$ (e.g., $\beta_0 = 0$)

for $k = 1$ to K **do**

Compute gradient of SSE(β):

$$\nabla f(\beta^{(t)}) = -\frac{1}{n} X^\top (y - X\beta^{(t)})$$

Gradient descent step:

$$\tilde{\beta}_k = \tilde{\beta}_k - \eta \nabla f(\beta_k)$$

Apply soft-thresholding (proximal operator for $\lambda\|\beta\|_1$):

$$\beta_{k+1}(j) = \text{sign}(\tilde{\beta}_k(j)) \cdot \max\Big(|\tilde{\beta}_k(j)| - \eta\lambda,\ 0\Big) \quad \text{for } j = 1, \cdots, p$$

Check stopping criterion:

$$\text{if } \|\beta_{k+1} - \beta_k\|_2 < \textbf{tol} \text{ then stop}$$

end for

return β_{k+1}

Problem 3:

Consider a lasso regression problem with five predictors and regularization parameter $\lambda = 0.5$. At iteration k, the gradient update with step size $\eta = 0.5$ produces the following values for the regression coefficients.

$$\begin{bmatrix} 2.1 \\ -3.5 \\ 0.2 \\ 1.3 \\ -0.5 \end{bmatrix}$$

What would the value of the regression coefficients be at iteration $k+1$?

Answer:

To obtain the updated coefficients $\beta^{(k+1)}$, we apply the soft-thresholding operator elementwise:

$$\beta_{k+1}(j) = \text{sign}(\tilde{\beta}_k(j)) \cdot \max\left(|\tilde{\beta}_k(j)| - \eta\lambda,\ 0\right), \quad j = 1, \cdots, 5$$

Since $\eta = 0.5$ and $\lambda = 0.5$, we compute $\eta\lambda = 0.25$. Next, apply the soft-thresholding to each component:

$$\begin{aligned}
\beta_{k+1}(1) &= \text{sign}(2.1) \cdot \max(2.1 - 0.25,\ 0) = 1 \cdot 1.85 = 1.85 \\
\beta_{k+1}(2) &= \text{sign}(-3.5) \cdot \max(3.5 - 0.25,\ 0) = -1 \cdot 3.25 = -3.25 \\
\beta_{k+1}(3) &= \text{sign}(0.2) \cdot \max(0.2 - 0.25,\ 0) = 1 \cdot 0 = 0 \\
\beta_{k+1}(4) &= \text{sign}(1.3) \cdot \max(1.3 - 0.25,\ 0) = 1 \cdot 1.05 = 1.05 \\
\beta_{k+1}(5) &= \text{sign}(-0.5) \cdot \max(0.5 - 0.25,\ 0) = -1 \cdot 0.25 = -0.25
\end{aligned}$$

Thus, the updated coefficients at iteration $k+1$ are:

$$\beta_{k+1} = \begin{bmatrix} 1.85 \\ -3.25 \\ 0 \\ 1.05 \\ -0.25 \end{bmatrix}$$

Problem 4: Consider the function

$$f(x, y) = x^2 + y^2 + \log(x) + \exp(y).$$

Use Newton's algorithm to perform **one iteration** starting from $(x_0, y_0) = (1, 0)$.

Answer:

Step 1: Compute the gradient

The gradient of f is:

$$\nabla f(x,y) = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} = \begin{bmatrix} 2x + \frac{1}{x} \\ 2y + \exp(y) \end{bmatrix}$$

At $(x_0, y_0) = (1, 0)$:

$$\nabla f(1,0) = \begin{bmatrix} 2 \cdot 1 + \frac{1}{1} \\ 2 \cdot 0 + \exp(0) \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

Step 2: Compute the Hessian

The Hessian matrix of f is:

$$H_f(x,y) = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} = \begin{bmatrix} 2 - \frac{1}{x^2} & 0 \\ 0 & 2 + \exp(y) \end{bmatrix}$$

At $(x_0, y_0) = (1, 0)$:

$$H_f(1,0) = \begin{bmatrix} 2-1 & 0 \\ 0 & 2+1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}$$

Step 3: Newton update

The Newton step is given by:

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} - H_f^{-1}(x_0, y_0) \cdot \nabla f(x_0, y_0)$$

Since the Hessian is diagonal, its inverse is easy to compute:

$$H_f^{-1}(1,0) = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{3} \end{bmatrix}$$

Thus,

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 3 \\ 1 \end{bmatrix} = \begin{bmatrix} 1-3 \\ 0-\frac{1}{3} \end{bmatrix} = \begin{bmatrix} -2 \\ -\frac{1}{3} \end{bmatrix}$$

Problem 5: Consider a test data set with the following responses $y = \{0, 1, 1, 0, 1, 1\}$.

A logistic regression model calculated the following predicted probabilities $\hat{y} = \{0.6, 0.9, 0.3, 0.2, 0.4, 0.75\}$. Calculate the confusion matrix if the threshold is set to $t = 0.5$.

How do your answers change if the threshold is set to $t = 0.3$.

Answer:

Consider the true responses:

$$y = \{0, 1, 1, 0, 1, 1\}$$

and the predicted probabilities:

$$\hat{y} = \{0.6, 0.9, 0.3, 0.2, 0.4, 0.75\}$$

Threshold $t = 0.5$

We convert probabilities to class predictions:

$$\hat{y}_{\text{class}} = \{1, 1, 0, 0, 0, 1\}$$

Compare with true labels:

- True Positives (TP): Predicted 1 and actual 1 = indices 2, 6 $\Rightarrow$ 2 cases
- False Positives (FP): Predicted 1 and actual 0 = index 1 $\Rightarrow$ 1 case
- True Negatives (TN): Predicted 0 and actual 0 = index 4 $\Rightarrow$ 1 case
- False Negatives (FN): Predicted 0 and actual 1 = index 3, 5 $\Rightarrow$ 2 cases

Confusion matrix at $t = 0.5$:

	Pred 0	Pred 1
True 0	1	1
True 1	2	2

Threshold $t = 0.3$

Convert probabilities:

$$\hat{y}_{\text{class}} = \{1, 1, 1, 0, 1, 1\}$$

Compare with true labels:

- TP: indices 2, 3, 5, 6 $\Rightarrow$ 4 cases
- FP: index 1 $\Rightarrow$ 1 case
- TN: index 4 $\Rightarrow$ 1 case
- FN: none

Confusion matrix at $t = 0.3$**:**

	Pred 0	Pred 1
True 0	1	1
True 1	0	4

2 Multiple choice Quiz Questions

Question 1: Which of the following best describes the ROC curve?

- ☐ A plot of precision vs. recall.
- ☑ A plot of true positive rate vs. false positive rate.
- ☐ A plot of sensitivity vs. specificity.
- ☐ A plot of true positives vs. false negatives.

Question 2: The Lasso regression is known for:

- ☐ Producing non-unique solutions.
- ☑ Selecting a subset of predictors.
- ☐ Being insensitive to regularization.
- ☐ Being equivalent to Ridge regression.

Question 3: What happens when the regularization parameter λ is very large in Ridge regression?

- ☐ The model overfits.
- ☐ The model becomes sparse.
- ☑ Regression coefficients shrink toward zero.
- ☐ AUC increases.

Question 4: The proximal gradient method is typically used when:

- ☐ The loss function is non-differentiable.
- ☑ The optimization problem has a composite structure, with one component being non-differentiable.

- ☐ Gradient descent is unstable.
- ☐ The loss function is quadratic.

Question 5: Which of the following is an example of a function with a simple proximal operator?

- ☑ ℓ_1 norm.
- ☐ Squared error loss.
- ☐ Binary cross-entropy loss.
- ☐ ℓ_0 norm.

Question 6: During k-fold cross-validation, test loss is calculated on:

- ☐ The entire data set.
- ☑ Only the training folds.
- ☐ Only the test fold.
- ☐ An entirely separate validation set.

Question 7: If validation loss starts increasing while training loss continues to decrease, this most likely indicates:

- ☐ Training requires more epochs.
- ☐ The step size used in the optimization algorithm is too large.
- ☑ Overfitting is occurring.
- ☐ Underfitting is occurring.

Question 8: For the Lasso problem, the proximal operator corresponds to:

- ☐ The identity operator.

- ☑ Soft thresholding.
- ☐ Hard thresholding.
- ☐ Projecting to a unit ball.

Question 9: Newton's algorithm for optimization uses which of the following in its update rule?

- ☐ Gradient only.
- ☐ Hessian only.
- ☑ Both gradient and Hessian.
- ☐ Neither gradient nor Hessian.

Question 10: A major computational challenge in implementing Newton's method is:

- ☐ Computing the gradient.
- ☑ Computing and inverting the Hessian matrix.
- ☐ Determining the step size.
- ☐ Computing the Hessian.

Question 11: When the Hessian is not positive definite, a common modification to Newton's algorithm is:

- ☐ To use gradient descent instead.
- ☑ To modify the Hessian to make it positive definite.
- ☐ To increase the step size.
- ☐ To use a different value for initialization.

Question 12: In the coordinate descent algorithm for linear regression, during each iteration:

- ☐ All regression coefficients are updated simultaneously.
- ☑ A single regression coefficient is updated while keeping others fixed.
- ☐ A random subset of regression coefficients is updated.
- ☐ All regression coefficients are updated, but in a specific order.

Question 13: Coordinate descent is most advantageous when:

- ☐ The dimension of the optimization problem is small.
- ☐ The optimization problem has multiple minima.
- ☐ The variables in the objective function have complex interactions.
- ☑ The optimization problem is high-dimensional and each coordinate-wise problem is easy to solve

Question 14: Newton's algorithm uses which second-order information?

- ☐ Gradient of the objective function only.
- ☑ Hessian matrix of the objective function.
- ☐ An identity matrix.
- ☐ Proximal operator.

Question 15: Which of the following is true about the AUC score?

- ☐ It is only used in regression problems.
- ☐ It can take negative values.
- ☑ A score closer to 1 indicates better classification.
- ☐ It measures calibration of predicted probabilities.

Question 16: The mathematical formulation of Lasso Regression adds which term to the ordinary least squares objective function?

- ☑ $\lambda \sum_{j=1}^{p} |\beta_j|$.
- ☐ $\lambda \sum_{j=1}^{p} \beta_j^2$.
- ☐ $\lambda \sum_{j=1}^{p} \beta_j$.
- ☐ $\lambda \max_j |\beta_j|$.

Question 17: In the proximal gradient algorithm, the proximal operator $\text{prox}_{t,g}(z)$ is defined as:

- ☑ $\arg\min_x g(x) + \frac{1}{2t}\|x - z\|_2^2$
- ☐ $\arg\min_x f(x) + \frac{1}{2t}\|x - z\|_2^2$
- ☐ $\arg\min_x g(x) + t\|x - z\|_2^2$
- ☐ $\arg\min_x f(x) + t\|x - z\|_2^2$

Question 18: Suppose that you have a data set comprising 1 million observations and 100,000 predictors that exhibit a high degree of multicollinearity. A computationally efficient algorithm to estimate the regression coefficients is:

- ☐ Gradient descent applied to the sum-of-squares errors loss function.
- ☐ Gradient descent applied to the sum-of-squares errors loss function, augmented with a regularization term that penalizes the sum of squared regression coefficients.
- ☐ Newton's algorithm applied to the sum-of-squares errors loss function, augmented with a regularization term that penalizes the sum of squared regression coefficients.
- ☑ Stochastic gradient descent applied augmented to the sum-of-squares errors loss function, augmented with a regularization term that penalizes the sum of squared regression coefficients.