Problem 1

Claim: $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}$ is invertible for $\lambda > 0$.

To prove this claim we can use eigendecomposition. Matrix $\mathbf{X}^T\mathbf{X}$ is symmetric, and we can apply eigendecomposition to it:

$$\mathbf{X}^T\mathbf{X} = \mathbf{Q}\Lambda\mathbf{Q}^T$$

where $\mathbf{Q}$ is the orthogonal matrix with the eigenvectors of $\mathbf{X}^T\mathbf{X}$

Λ is the diagonal matrix with the eigenvalues of $\mathbf{X}^T\mathbf{X}$

$\mathbf{X}^T\mathbf{X} = \mathbf{Q}\Lambda\mathbf{Q}^T$ is a square matrix. If we can show all the eqigenvalues of this matrix are strictly greater than 0, we can prove the claim is true, which $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}$ is invertible for $\lambda > 0$.

Eigenvalues of $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}$ are the diagonals of $\Lambda + \lambda\mathbf{I}$. Since the all elements of Λ are greater than or equal to 0, all the eigenvalues of $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}$ are strictly greater than 0 because $\lambda > 0$. $\square$

Problem 2.a

From the problem definition, the expected loss of a prediction function $f(x)$ in modeling y using loss function $\ell(f(x), y)$ is given by;

$$\mathbb{E}_{(x,y)}[\ell(f(x), y)] = \int_x \int_y \ell(f(x), y)p(x, y)\, dy\, dx = \int_x \left\{ \int_y \ell(f(x), y)p(y|x)dy \right\} p(x)dx$$

We take the gradient of the given expected loss function with respect to f and make it equal to 0 to find the optimal $f(x)$:

$$\frac{\partial}{\partial f}\mathbb{E}_{(x,y)}[\ell(f(x), y)] = 0$$

$$= \int_x \frac{\partial}{\partial f}\left\{ \int_y (f(x) - y)^2 p(y|x)dy \right\} p(x)dx = 0$$

$$= \int_x 2\left\{ \int_y (f(x) - y)p(y|x)dy \right\} p(x)dx = 0,$$

we can eliminate 2 from this equality.

Assign optimal $f(x)$ *as* $f^*(x)$

$$\implies \int_x \left\{ \int_y (f^*(x) - y)p(y|x)dy \right\} p(x)dx = 0$$

$$\implies \int_y (f^*(x) - y)p(y|x)dy = 0$$

$$\implies \int_y f^*(x)p(y|x)dy = \int_y yp(y|x)dy$$

$$\implies f^*(x) = \mathbb{E}[y|x].$$

The optimal prediction function $f^*(x) = \mathbb{E}[y|x]$ means that, when using the loss function $\ell(f(x), y) = (f(x) - y)^2$, estimating y as the expected value of y given x is the best way in modeling the variable y which is our target.

Problem 2.b
Use $\ell(f(x), y) = |f(x) - y|$.

$$\mathbb{E}_{(x,y)}[\ell(f(x), y)] = \int_x \left\{ \int_y |f(x) - y| p(y|x) dy \right\} p(x) dx = 0$$

Following this,

$$\arg\min_f \mathbb{E}_{(x,y)}[|f(x) - y|] \qquad \text{will yield optimum value for } f(x),\ f^*(x)$$

Consider two different cases: $f(x) \geq y$ and $f(x) < y$

When $f(x) \geq y$:

$$\mathbb{E}_{(x,y)}[|f(x) - y|] = \int_x \left\{ \int_y (f(x) - y) p(y|x) dy \right\} p(x) dx = 0$$

When $f(x) < y$:

$$\mathbb{E}_{(x,y)}[|f(x) - y|] = \int_x \left\{ \int_y (y - f(x)) p(y|x) dy \right\} p(x) dx = 0$$

Taking the gradient of $\mathbb{E}_{(x,y)}[|f(x) - y|]$ with respect to $f(x)$ and equating it to 0 will give the optimal prediction, $f^*(x)$:

$$\int_x \left\{ \int_y sgn(f(x) - y) p(y|x) dy \right\} p(x) dx = 0$$

$$\implies f^*(x) = median(y|x).$$

The optimal prediction $f^*(x) = median(y|x)$ means, when using the loss function $\ell(f(x), y) = |f(x) - y|$, estimating y as the conditional median of y given x is the best way in modeling he variable y.

Problem 3
An output example is the following:

```
Ridge regression CV MSE values ['0.52182', '0.52482', '0.53391', '0.56698', '0.55272',
'0.47963', '0.52045', '0.53880', '0.49403', '0.51004', 'Mean: 0.52432  Std: 0.02463']
Logistic Regression CV error rates ['0.01786', '0.07143', '0.01786', '0.00000', '0.10714',
'0.08929', '0.07143', '0.07143', '0.03571', '0.05357', 'Mean: 0.05357  Std: 0.03293']
```

Please see `my_cross_val.py` for details.

Problem 4

MSE for Ridge Regression $\lambda = 0.01$											
F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Mean	SD
0.50053	0.52606	0.54548	0.55008	0.49656	0.53474	0.48806	0.51549	0.58072	0.50555	0.52433	0.03177

MSE for Ridge Regression $\lambda = 0.1$											
F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Mean	SD
0.46847	0.57921	0.55187	0.50076	0.55503	0.48657	0.48536	0.56731	0.56390	0.48940	0.52479	0.01838

MSE for Ridge Regression $\lambda = 1$											
F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Mean	SD
0.57144	0.52921	0.56122	0.51490	0.51280	0.55965	0.55227	0.49218	0.53942	0.57396	0.54070	0.01611

MSE for Ridge Regression $\lambda = 10$											
F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Mean	SD
0.59823	0.56545	0.60869	0.58888	0.57241	0.61952	0.56662	0.63420	0.56114	0.56391	0.58790	0.02428

MSE for Ridge Regression $\lambda = 100$											
F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Mean	SD
0.59391	0.60637	0.57598	0.60588	0.60146	0.61539	0.57209	0.61453	0.63485	0.60729	0.60278	0.02812

MSE for Lasso Regression $\lambda = 0.01$											
F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Mean	SD
0.55108	0.50055	0.51191	0.50764	0.62514	0.52050	0.52077	0.52644	0.54068	0.48929	0.52940	0.03626

MSE for Lasso Regression $\lambda = 0.1$											
F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Mean	SD
0.60795	0.58397	0.63233	0.60449	0.61354	0.62158	0.54487	0.62279	0.60334	0.62015	0.60550	0.02389

MSE for Lasso Regression $\lambda = 1$											
F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Mean	SD
0.93240	0.94889	0.96051	0.98593	0.97911	0.89187	0.92342	0.93550	0.99874	0.96073	0.95171	0.03053

MSE for Lasso Regression $\lambda = 10$											
F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Mean	SD
1.38949	1.33062	1.32248	1.36551	1.25132	1.33459	1.31729	1.28809	1.34931	1.35949	1.33082	0.03771

MSE for Lasso Regression $\lambda = 100$											
F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Mean	SD
1.23668	1.30486	1.33603	1.36577	1.28324	1.34158	1.37368	1.27391	1.39811	1.40165	1.33155	0.05271

As the λ increases mean of the folds increases for both MSE types. Lower λ means more optimal. However, note that, too small λ could result in overfitting. Larger λ means more regularization and it makes the model less effective at fitting the data.
$\lambda = 0.01$ is the most optimal for both methods.

MSE for Ridge Regression on the Test Data $\lambda = 0.01$											
F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Mean	SD
0.52205	0.42542	0.46055	0.48274	0.46784	0.53433	0.49746	0.50033	0.44498	0.46261	0.47983	0.03237

MSE for Lasso Regression on the Test Data $\lambda = 0.01$											
F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Mean	SD
0.56778	0.50461	0.45643	0.49001	0.44344	0.50803	0.59268	0.43418	0.44206	0.41352	0.48527	0.05609

Overall, Ridge Regression gives better results than Lasso as the corresponding fold values are smaller and therefore their means are smaller compared to Lasso.

Problem 5

Based on the projected data points and trial-and-error, optimal values of lambda lie between [-0.08,0]. Therefore, you'll only see $\lambda \in$ [-0.008, -0.007, -0.006, -0.005, -0.004, -0.003, -0.002, -0.001, 0] in `hw1_q5.py`.
In general, as $|\lambda|$ increases mean of the error folds also increases. However, this is true until the optimal λ is reached. Beyond the optimal λ, model starts to overfit therefore mean increases. Among the experienced λ values, -0.002 gives the best results for the error folds as it yields the least mean. Therefore, this optimal λ is used on the test data. As you can see, the mean of the folds is the least, smaller than all the means found using several $\lambda's$. This shows that training was successful.
Please see the tables below.

Error Rates for LDA $\lambda = -0.008$											
F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Mean	SD
0.09000	0.05500	0.07000	0.09000	0.12500	0.09000	0.07000	0.06500	0.05500	0.07500	0.07850	0.02001

Error Rates for LDA $\lambda = -0.007$											
F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Mean	SD
0.03500	0.05500	0.06000	0.04500	0.04500	0.03500	0.05000	0.08500	0.03000	0.05000	0.04900	0.01497

Error Rates for LDA $\lambda = -0.006$											
F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Mean	SD
0.04000	0.02000	0.02500	0.04000	0.02000	0.03500	0.04000	0.02500	0.04000	0.00000	0.02850	0.01246

Error Rates for LDA $\lambda = -0.005$											
F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Mean	SD
0.00500	0.03000	0.01000	0.00000	0.03500	0.03500	0.01500	0.00500	0.02500	0.03000	0.01900	0.01281

Error Rates for LDA $\lambda = -0.004$											
F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Mean	SD
0.01000	0.02500	0.00500	0.01000	0.01000	0.02500	0.01000	0.00500	0.01500	0.02000	0.01350	0.00709

Error Rates for LDA $\lambda = -0.003$											
F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Mean	SD
0.01000	0.01000	0.00000	0.00000	0.01500	0.02000	0.01500	0.01000	0.02000	0.02000	0.01200	0.00714

Error Rates for LDA $\lambda = -0.002$											
F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Mean	SD
0.01000	0.01000	0.01000	0.02000	0.01500	0.00000	0.01500	0.01000	0.00500	0.01000	0.01050	0.00522

Error Rates for LDA $\lambda = -0.001$											
F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Mean	SD
0.01000	0.00000	0.01000	0.01500	0.03000	0.01000	0.00500	0.01500	0.02500	0.03500	0.01550	0.01059

Error Rates for LDA $\lambda = 0$											
F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Mean	SD
0.02000	0.00000	0.03500	0.02000	0.02500	0.01000	0.02500	0.05000	0.02500	0.03000	0.02400	0.01281

Error Rates for LDA on the Test Data $\lambda = -0.002$											
F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Mean	SD
0.02500	0.02500	0.00000	0.00000	0.00000	0.00000	0.02500	0.02500	0.00000	0.00000	0.01000	0.01225