EECS289A Introduction to Machine Learning, Project T Final - Quiz Questions

Han Liu, Peng Tan, Dilu Xu, Jinyan Zhao

Department of Civil Engineering

Universitu of California, Berkeley

Berkeley, CA 94704

{han_liu, tanpeng, diluxu, jinyan_zhao}@berkeley.edu

December 10, 2020

1 Lagrange Duality and the LASSO

This question is extra credit!

It is often the case that we will wish to include additional constraints on our regularization terms. For instance if we wish to limit the size of our weights to be below a specific constant term, or if we wish that all of our weights should be non-negative, etc. in these scenarios, Lagrange multipliers are a powerful tool that allows us to incorporate these constraints into our objective functions. Also, in general, we have seen for $n \times p$ data, with n observation and p features that $n >> p$ where p is the dominant factor for the computation time in things like linear regression training, which is fine. However, in some scenarios $p >> n$, and this becomes a problem such as when we use the kernel trick! We find that reformulating these problems into their duals gives us the alternative dominant factor.

Here we will explore both of these phenomena simultaneously and see how to formulate a problem with the Lagrangian dual!

This problem follows the formulation in Osborne et al, 2000:

http://www4.stat.ncsu.edu/ hzhang/paper/osborne.pdf

By its original definition, the LASSO estimator solves the optimization problem

$$\text{minimize}_{w_1, \cdots, w_m} f(\boldsymbol{w}) = \frac{1}{2} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{m} x_{ij} w_j)^2 \tag{1}$$

subject to

$$f_1(\boldsymbol{w}) = \sum_{j=1}^{m} |w_j| \leq t \quad \text{for some } t > 0 \tag{2}$$

I.e. it minimizes half the sum of squared errors, subject to the constraint that $|\hat{w}|_1 \leq t$.

In this question, you will use Lagrangian duality to show that this formulation is equivalent to the following optimization problem (i.e. standard L1-regularized regression):

$$\text{minimize}_{w_1,\cdots,w_m} f(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{m}x_{ij}w_j)^2 + \lambda\sum_{j=1}^{m}|w_j| \quad \text{for some } \lambda > 0 \tag{3}$$

Specifically: you will convert the constrained optimization problem 1 into a Lagrange Multiplier problem. From there you will find that when w is an optimal solution to the constrained optimization problem and λ is the Lagrange multiplier associated with w, then w is also an optimal solution of the regularized regression problem with penalty coefficient λ. Your task is to complete part of this proof.

1. [5 points] The function 3 is convex in w (quadratic plus L1 norm), which means its minima satisfy ∇w(Eq. 3) = 0. Compute the gradient with respect to w of this expression and show that the resulting condition for any optimal w is

$$0 = -X^T(y - XW) + \lambda v \tag{4}$$

where $v = (v_1, \cdots, v_m) \in \mathbb{R}^m$ satisfies $v_i = \begin{cases} 1 & \text{if } w_i > 0 \\ -1 & \text{if } w_i < 0 \\ \in [-1,1] & \text{if } w_i = 0 \end{cases}$ This awkward-seeming vector

v is the result of taking the derivative of absolute value terms $|w_i|$, where the wi are of unknown sign. *Hint: this is similar to the derivation of the normal equation in the regression notes.*

Answer:

The same as the derivation of the normal equation in the notes; no need to provide much more detail. The v term is new and arises directly from taking partial derivatives of the $|w_i|$ terms.

2. [5 points] The next step is to take each case of the KKT conditions and solve the constrained optimization problem using Lagrange multipliers. To do this, note that if X is full-rank and $m \le n$, then the ordinary least-squares regression (OLS) problem has a unique solution:

$$\hat{w} = (X^TX)^{-1}X^TY \tag{5}$$

On the other hand, if $m > n$, then the solution $\hat{w}$ to the OLS problem is not unique. Let W be the set of solutions to a particular OLS problem. One can show (see Osbourne et al.) that if t_0 is defined as

$$t_0 = \min_{\boldsymbol{w} \in W} ||\boldsymbol{w}||_1 \quad \text{L1 norm} \tag{6}$$

then the constraint 2 is active if and only if $t < t_0$. In other words, if $t \ge t_0$, then the constraint doesn't matter– a solution to 1 just has to optimize the unconstrained OLS problem. On the other hand, if $t < t_0$, then the constraint 2 must be active for any solution w; in other words, this inequality constraint becomes an equality. Moreover, one can show that when $t < t_0$, at least one optimal w for the constrained problem 1 is guaranteed to exist on the boundary of the feasible region (the feasible region is the set of all w such that the constraint 2 is satisfied).

Based on this discussion, we only need to find solutions for one case of the KKT conditions when $t < t_0$. Which case is this, and why?

Answer:

When t < t0, the constraint must be active at any optimum. This means we must have $t - f_1(w) = 0$.

3. [10 points] Define
$$L^*(\boldsymbol{w}) = \max_{\lambda \geq 0} L(\boldsymbol{w}, \lambda) \tag{7}$$

Show that
$$L^*(\boldsymbol{w}) = \begin{cases} f(\boldsymbol{w}) & \text{if } t - f_1(w) \geq 0 \\ \infty & \text{if } t - f_1(w) < 0 \end{cases} \tag{8}$$

Argue that, therefore, when $t < t_0$, minimizing $L^*(\boldsymbol{w})$ over all possible w is equivalent to solving the constrained optimization problem 1. Minimizing $L^*(\boldsymbol{w})$ is known as the primal problem. *Hint: Show that if w optimizes the constrained problem, then it also must minimize L^*. Next, show that if w minimizes L^*, then it also must optimize the constrained problem. Proof by contradiction may be useful.*

Answer:

In the first case: when $t - f_1(w) \geq 0$, then $L(w, \lambda)$ is maximized when $\lambda = 0$. This leaves only $f(w)$. In the second case: when $t - f_1(w) < 0$, then bigger λ always makes $L(w, \lambda)$ bigger; hence the max goes to infinity.

Suppose w minimizes the constrained optimization problem. Then from the previous question, we must have $t - f_1(w) = 0$. Suppose there exists w_0 such that $L^*(\boldsymbol{w}_0) < L^*(\boldsymbol{w})$. Since $t - f_1(w) = 0$, we must then have $f(\boldsymbol{w}_0) < f(\boldsymbol{w})$ with the constraint satisfied. But then w does not minimize the constrained optimization problem; contradiction. On the other hand, suppose w minimizes L^*. For any $t \geq 0$, there always exists some w with L1 norm not greater than t. Hence w is in the feasible region $t - f1(w) \geq 0$, and the minimal $L^*(\boldsymbol{w})$ is finite. Suppose $f(w0) < f(w)$ for feasible $w0$. Then we would also have $L^*(\boldsymbol{w}_0) < L^*(\boldsymbol{w})$, contradition. This completes the proof of equivalence between the two problems.

4. [5 points] Define
$$g(\lambda) = \min_{\boldsymbol{w} \in \mathbb{R}^m} L(\boldsymbol{w}, \lambda) \tag{9}$$

Maximizing $g(\lambda)$ for $\lambda \geq 0$ is the dual problem to the primal shown above.

One can show that, for fixed λ, $L(\boldsymbol{w}, \lambda)$ has at least one minimum over w. We can solve for a minimum in the usual way: by taking the gradient of L with respect to w and setting it to zero. Do this to show that if $\bar{w}$ minimizes $L(\boldsymbol{w}, \lambda)$ for fixed λ, then

$$0 = -X^T(y - X\bar{w}) + \lambda v \tag{10}$$

with v defined as before, in 4.

Answer:

Same as before; only the extra $-\lambda t$ term is new, and it vanishes when we take the gradient with respect to $\boldsymbol{w}$.

5. [5 points] Show that $\boldsymbol{v}^T\bar{w} = ||\bar{w}||_1$. Consequently, show that if $\bar{w}$ minimizes $L(\boldsymbol{w}, \lambda)$ for fixed Λ, then

$$\lambda = \frac{(y - X\bar{w})^T X\bar{w}}{||\bar{w}||_1} \tag{11}$$

We call this value of λ the Lagrange multiplier associated with $\bar{w}$. *Hint: recall that $(AB)^T = B^T A^T$.*

Answer:

First, $\boldsymbol{v}^T \bar{w} = \sum_{i=1}^m v_i \bar{w}_i$. By definition, each v_i has magnitude 1 and the same sign as $\bar{w}_i$. Hence their product is $|\bar{w}_i|$. The sum of these is the L1 norm of $\boldsymbol{w}$.

Next, we start with

$$0 = -X^T(y - X\bar{w}) + \lambda v, \quad \text{and so} \tag{12}$$

$$\lambda v = X^T(y - X\bar{w}), \quad \text{and so} \tag{13}$$

$$\lambda v^T = (y - X\bar{w})^T X, \quad \text{and so} \tag{14}$$

$$\lambda v^T \bar{w} = (y - X\bar{w})^T X\bar{w}, \quad \text{and so} \tag{15}$$

$$\lambda = \frac{(y - X\bar{w})^T X\bar{w}}{||\bar{w}||_1} \quad \text{as desired.} \tag{16}$$

It turns out that strong duality holds for this problem (see Theorem 3 in Osborne et al), which means: if w^* is a solution of the primal problem (equivalently, of the constrained optimization problem 1) and λ^* is the Lagrange multiplier corresponding to w^*, then λ^* is a solution of the dual problem, and hence the optimal primal and dual function values are equal:

$$g(\lambda^*) = L^*(\boldsymbol{w}^*) = f(\boldsymbol{w}^*) \tag{17}$$

for f as defined in 1.

Let's review: we began by trying to solve the constrained optimization problem 1. We saw that this is equivalent to solving the primal problem 7. Due to strong duality, we can solve either the primal or dual problem to obtain an optimal w for 1. Given a solution w for 1, we can compute the Lagrange multiplier associated with it using 11. On the other hand, given λ, any optimal w for the penalized regression problem 3 satisfies exactly the same equations (4) as are satisfied by solutions to 1. This means: when w is an optimal solution to the constrained optimization problem and λ is the Lagrange multiplier associated with w, then w is also an optimal solution of the penalized regression problem with penalty coefficient λ.

2 Basic Question

1. A computer program is said to learn from experience E with respect to some task T and some performance measure P if its performance on T, as measured by P, improves with experience E. Suppose we feed a learning algorithm a lot of historical weather data, and have it learn to predict weather. What would be a reasonable choice for P?
   a. The probability of it correctly predicting a future date's weather.
   b. The weather prediction task.
   c. The process of the algorithm examining a large amount of historical weather data.
   d. None of these.
   Answer:
   a

2. A computer program is said to learn from experience E with respect to some task T and some performance measure P if its performance on T, as measured by P, improves with experience E. Suppose we feed a learning algorithm a lot of historical weather data, and have it learn to predict weather. In this setting, what is T?
   a. The weather prediction task.
   b. The probability of it correctly predicting a future date's weather.
   c. The process of the algorithm examining a large amount of historical weather data.
   d. None of these.

   Answer:
   a

3. Suppose you are working on weather prediction, and use a learning algorithm to predict tomorrow's temperature (in degrees Centigrade/Fahrenheit).
   Would you treat this as a classification or a regression problem?
   a. Regression
   b. Classification

   Answer:
   a

4. Suppose you are working on weather prediction, and your weather station makes one of three predictions for each day's weather: Sunny, Cloudy or Rainy. You'd like to use a learning algorithm to predict tomorrow's weather.
   Would you treat this as a classification or a regression problem?
   a. Regression
   b. Classification

   Answer:
   b

3   Elastic Net

In class we discussed two types of regularization, l_1 and l_2. Both are useful, and sometimes it is helpful combine them, giving the objective function below (**in this problem we have excluded w_0 for simplicity**):

$$F(\boldsymbol{w}) = \frac{1}{2}\sum_{j=1}^{n}(y^{(j)} - \sum_{i=1}^{d}\boldsymbol{w}_i\boldsymbol{x}_i^{(j)})^2 + \alpha\sum_{i=1}^{d}|\boldsymbol{w}_i| + \frac{\lambda}{2} + \sum_{i=1}^{d}\boldsymbol{w}_i^2 \tag{18}$$

Here, $(\boldsymbol{x}^{(j)}, y^{(j)})$ is j-th example in the training data, w is a d dimensional weight vector, λ is a regularization parameter for the l_2 norm of w, and α is a regularization parameter for the l_1 norm of w. This approach is called the Elastic Net, and you can see that it is a generalization of Ridge and Lasso regression: It reverts to Lasso when $\lambda = 0$, and it reverts to Ridge when $\alpha = 0$. In this question, we are going to derive the coordinate descent (CD) update rule for this objective.

Let g, h, c be real constants, and consider the function of x

$$f_1(x) = c + gx + \frac{1}{2}hx^2 \tag{19}$$

1. [4 points] What is the x^* that minimizes $f_1(x)$? (i.e. calculate $x^* = \arg\min f_1(x)$)
   Answer:
   Take the gradient of $f_1(x)$ and set it to 0.

   $$g + hx = 0 \tag{20}$$

   $$x^* = -\frac{g}{h} \tag{21}$$

   Let α be an additional real constant, and consider another function of x

   $$f_2(x) = c + gx + \frac{1}{2}hx^2 + \alpha|x|(h > 0, \alpha > 0) \tag{22}$$

   This is a piecewise function, composed of two quadratic functions:

   $$f_2^-(x) = c + gx + \frac{1}{2}hx^2 - \alpha x \tag{23}$$

   and

   $$f_2^+(x) = c + gx + \frac{1}{2}hx^2 + \alpha x \tag{24}$$

   Let $\tilde{x}^- = \arg\min_{x\in\mathbb{R}} f_2^-(x)$ and $\tilde{x}^+ = \arg\min_{x\in\mathbb{R}} f_2^+(x)$.
   (**Note:** The argmin is taken over $(-\infty, +\infty)$ here.

2. [6 points] What are $\tilde{x}^+$ and $\tilde{x}^+$? Show that $\tilde{x}^- \geq \tilde{x}^+$.
   Answer:
   Using the answer from part 1), we get $\tilde{x}^+ = -\frac{g+\alpha}{h}$ and $\tilde{x}^- = -\frac{g-\alpha}{h}$. Since $\tilde{x}^- - \tilde{x}^+ = \frac{2\alpha}{h} \geq 0$, we have $\tilde{x}^- \geq \tilde{x}^+$.

3. [12 points] Draw a picture of $f_2(x)$ in each of the three cases below:
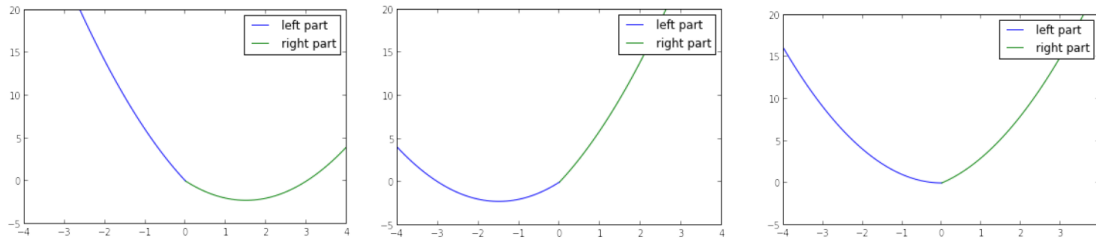
(a) $\tilde{x}^- > 0$, $\tilde{x}^+ > 0$

(b) $\tilde{x}^- < 0$, $\tilde{x}^+ < 0$

(c) $\tilde{x}^- > 0$, $\tilde{x}^+ < 0$

For each case, mark the minimum as either 0, $\tilde{x}^-$, or $\tilde{x}^+$. (You do not need to draw perfect curves, just get the rough shape and the relative locations of the minima to the x-axis)

Answer:

Fig. 1 gives the example picture of three cases. To understand the answer, note the following:

- $f_2^+(0) = f_2^-(0)$, this means the two piece of f_2 match each other at x $= 0$.

- When $\tilde{x}^- \geq 0$, the minimum of the left part is at 0, i.e. $\arg\min_{x \in (-\infty, 0]} f_2^-(x) = 0$

- When $\tilde{x}^- \leq 0$, the minimum of the left part is at $\tilde{x}^-$, i.e. $\arg\min_{x \in (-\infty, 0]} f_2^-(x) = \tilde{x}^-$

- Similar rules holds for the right part of the curve.

(a) $\tilde{x}^- > 0$, $\tilde{x}^+ > 0$, the minima is at $\tilde{x}^+$.

(b) $\tilde{x}^- < 0$, $\tilde{x}^+ < 0$, the minima is at $\tilde{x}^-$.

(c) $\tilde{x}^- > 0$, $\tilde{x}^+ < 0$, the minima is at 0.

Fig. 1. Example picture of three cases.

4 Elastic Net Plus

1. [8 points] Write x^*, the minimum of $f_2(x)$, as a piecewise function of g. (Hint: what is g in the cases above?)
   Answer:
   Summarizing the result from the previous question, we have

$$
x^* = \begin{cases} \tilde{x}^+ & \text{if } \tilde{x}^+ > 0 \\ \tilde{x}^- & \text{if } \tilde{x}^- < 0 \\ 0 & \text{if } \tilde{x}^+ < 0, \tilde{x}^- > 0 \end{cases} \tag{25}
$$

This is equivalent to the soft-threshold function

$$
x^* = \begin{cases} -\frac{g+\alpha}{h} & \text{if } g < -\alpha \\ 0 & \text{if } g \in [-\alpha, \alpha] \\ -\frac{g-\alpha}{h} & \text{if } g > \alpha \end{cases} \tag{26}
$$

2. [8 points] Now let's derive the update rule for w_k. Fixing all parameters but w_k, express the objective function in the form of Eq. (5) (here w_k is our x). You do not need to explicity write constant factors, you can put them all together as c. What are g and h? Then putting it all together, write down the coordinate descent algorithm for the Elastic Net (**again excluding w_0**). You can write the update for w_k in terms of the g and h you calculated in the previous question.
   Answer:

$$
g = \sum_{j=1}^{n} x_k^{(j)} \left(\sum_{i \neq k} \boldsymbol{w}_i \boldsymbol{x}_i^{(j)} - y_i \right) \tag{27}
$$

$$
h = \lambda + \sum_{j=1}^{n} (x^{(j)})^2 \tag{28}
$$

The algorithm is shown in Algorithm. 1. You can compare it to the CD algorithm for Lasso, the only difference is adding λ to h.

Algorithm 1: Coordinate Descent for Elastic Net

---
while *not converged* **do**
  for $k \in \{1, 2, \cdots, d\}$ **do**
   $g = \sum_{j=1}^{n} x_k^{(j)} (\sum_{i \neq k} \boldsymbol{w}_i \boldsymbol{x}_i^{(j)} - y_i)$;
   $h = \lambda + \sum_{j=1}^{n} (x^{(j)})^2$;
   $w_k = \begin{cases} -\frac{g+\alpha}{h} & \text{if } g < -\alpha \\ 0 & \text{if } g \in [-\alpha, \alpha] \\ -\frac{g-\alpha}{h} & \text{if } g > \alpha \end{cases}$;
  end
end

---

5   True or False questions

Answer True or False. Justify your answer very briefly in 1-2 sentences.

1. Increasing the regularization parameter λ in lasso regression leads to sparser regression coefficients.
   Answer:
   True. Larger regularization parameter penalizes non-zero coefficients more, leading to sparser solution.

2. In Lasso regression, can the regulariser increase the sparsity of the resulting solutions?
   Answer:
   True, Lasso introduces the L1 norm penalty, so increasing the penalty will cause more and more of the parameters to be driven to zero. As seen in Fig. 2, where the red contours are representing the error function and the solid blue areas the constraint regions for the parameters, in the Lasso case, the function is likely to hit the constraint region on one of the four corners. This will make a parameter equal with zero.
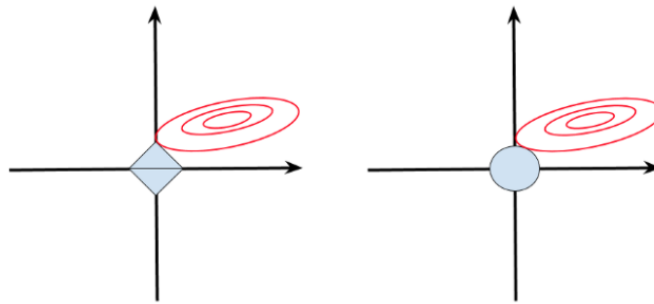

Fig. 2. Spam-Estimation picture for the LASSO (left) and Ridge regression (right).

3. Is the objective of Lasso regression differentiable and convex?
   Answer:
   False, it is not differentiable. As Lasso uses L1 norm penalty, this L1 loss function is defined based on the absolute value of the difference between two values. Since the loss function is a modulus function, there will exist an inflection point where the slope cannot be computed and the function is not differentiable. In contrast, for the case of Ridge regression, where we use L2 norm penalty, we can get the slope at any point and the objective is differentiable.

4. Are the objectives of Ridge and Lasso regression both convex and have closed-form solutions?
   Answer:
   False, they are both convex, but Lasso does not always have a closed-form solution. LASSO regression must be solved numerically by using quadratic programming or more general convex optimisation methods.

5. Is it true that the L1 term in Lasso has the following purposes: performing feature selection, compensating for overfitting and smoothing?
   Answer:

False. The first two purposes are true, but not the last one. As it is mentioned early, both L1 and L2 terms are introduced to compensate for overfitting. By making the parameters equal with zero, the L1 norm discards some features and, thus, performs feature selection. In terms of smoothing, we cannot say L1 has this purpose, as the L1 loss function does not have continuous derivatives.

6   Quick questions

Explain **in one or two sentences** why the statements are true (or false).

1. L2 loss is more robust to outliers than L1 loss.
   Answer:
   False
   The gradient of L2 loss can grow without bound whereas the L1 loss gradient is bounded, hence the influence of an outlier is limited.

2. Logistic loss is better than L2 loss in classification tasks.
   Answer:
   True
   With logistic loss, correctly classified points that are far away from the decision boundary have much less impact on the decision boundary.

3. In terms of feature selection, L2 regularization is preferred since it comes up with sparse solutions.
   Answer:
   False
   L1 regularization (LASSO) comes up with sparse solutions due to nonvanishing gradient at 0.

7 Gradient descent

Denote by $\boldsymbol{x} \in \mathbb{R}^d$ data, by $\boldsymbol{w} \in \mathbb{R}^d$ the weight vector, by $y \in \mathbb{R}y$ lables, by $\lambda > 0$ a regularization constant, and by m the total number of data points. Let R(w) be the regularized risk function:

$$R(w) := \frac{1}{m} \sum_{i=1}^{m} l(y_i - \langle w, x_i \rangle) + \frac{\lambda}{2} ||\boldsymbol{w}||^2 \quad \text{where} \quad l(\xi) = \begin{cases} \frac{1}{2}\xi^2 & \text{if } |\xi| < 1 \\ |\xi| - \frac{1}{2} & \text{otherwise} \end{cases} \tag{29}$$

1. Calculate the batch gradient of R(w) with respect to w.
   Answer:
   Define g_i as the gradient of ith data point.

   $$g_i = \begin{cases} (\langle w, x_i \rangle - y)x_i & \text{if} | \langle w, x_i \rangle - y| < 1 \\ \text{sgn}(\langle w, x_i \rangle - y)x_i & \text{otherwise} \end{cases} \tag{30}$$

   $$\frac{\partial R}{\partial w} = \frac{1}{m} \sum_{i=1}^{m} g_i + \lambda w \tag{31}$$

2. Write out a stochastic gradient descent learning algorithm for minimizing R(w).
   Answer:
   Randomly select a training instance x_i, update w as follows:

   $$\boldsymbol{w} \leftarrow (1 - \lambda\eta_t)\boldsymbol{w} - \eta_t g_i \tag{32}$$

3. Name two common choices for choosing the learning rate and briefly explain their properties.
   Answer:
   $O(\frac{1}{\sqrt{t}}), O(\frac{1}{t})$ for strong convexity, Polynomial decay, AdaGrad, etc.

8 Logistic Regression

1. For logistic regression, the gradient is given by $\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^i) x_j^{(i)}$. Which of these is a correct gradient descent update for logistic regression with a learning rate of α? Check all that apply.

   a. $\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)-y^i}) x_j^{(i)}$ (simultaneously update for all j).

   b. $\theta := \theta - \alpha \frac{1}{m} \sum_{i=1}^{m} (\theta^T x - y^{(i)}) x^{(i)}$.

   c. $\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(\frac{1}{1+e^{-\theta^T x^{(i)}}} - y^{(i)} \right) x_j^{(i)}$ (simultaneously update for all j).

   d. $\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)-y^i}) x^{(i)}$ (simultaneously update for all j).

   Answer:
   ac

2. Suppose that you have trained a logistic regression classifier, and it outputs on a new example a prediction $h_\theta(x) = 0.2$. This means (check all that apply):

   a. Our estimate for $P(y = 1|x; \theta)$ is 0.8.
   b. Our estimate for $P(y = 0|x; \theta)$ is 0.8.
   c. Our estimate for $P(y = 1|x; \theta)$ is 0.2.
   d. Our estimate for $P(y = 0|x; \theta)$ is 0.2.

   Answer:
   bc

9 Multiple choice I

In this part, you only need to choose one correct answer.

1. You run gradient descent for 15 iterations with $\alpha = 0.3$ and compute after each iteration. You find that the value of $J(\theta)$ decreases slowly and is still decreasing after 15 iterations. Based on this, which of the following conclusions seems most plausible?
   a. Rather than use the current value of α, it'd be more promising to try a larger value of α (say $\alpha = 1.0$).
   b. Rather than use the current value of α, it'd be more promising to try a smaller value of α (say $\alpha = 0.1$).
   c. $\alpha = 0.3$ is an effective choice of learning rate.
   Answer:
   a

2. You run gradient descent for 15 iterations with $\alpha = 0.3$ and compute after each iteration. You find that the value of $J(\theta)$ decreases quickly then levels off. Based on this, which of the following conclusions seems most plausible?
   a. Rather than use the current value of α, it'd be more promising to try a larger value of α (say $\alpha = 1.0$).
   b. Rather than use the current value of α, it'd be more promising to try a smaller value of α (say $\alpha = 0.1$).
   c. $\alpha = 0.3$ is an effective choice of learning rate.
   Answer:
   c

3. Suppose you have $m = 23$ training examples with $n = 5$ features (excluding the additional all-ones feature for the intercept term, which you should add). The normal equation is $\theta = (X^T X)^{-1} X^T y$. For the given values of m and n, what are the dimensions of θ, X, and y in this equation?
   a. X is 23×5, y is 23×1, θ is 5×5
   b. X is 23×6, y is 23×6, θ is 6×6
   c. X is 23×6, y is 23×1, θ is 6×1
   d. X is 23×5, y is 23×1, θ is 5×1
   Answer:
   c
   X has m rows and n+1 columns (+1 because of the $x_0 = 1$ term). y is m-vector. θ is an $(n + 1)$-vector

4. Suppose you have a dataset with $m = 1000000$ examples and $n = 200000$ features for each example. You want to use multivariate linear regression to fit the parameters θ to our data. Should you prefer gradient descent or the normal equation?
   a. Gradient descent, since it will always converge to the optimal θ.
   b. Gradient descent, since $(X^T X)^{-1}$ will be very slow to compute in the normal equation.
   c. The normal equation, since it provides an efficient way to directly find the solution.
   d. The normal equation, since gradient descent might be unable to find the optimal θ.
   Answer:
   b
   With $n = 200000$ features, you will have to invert a $200001x200001$ matrix to compute the normal

14

equation. Inverting such a large matrix is computationally expensive, so gradient descent is a good choice.

5. Which of the following are reasons for using feature scaling?
a. It is necessary to prevent gradient descent from getting stuck in local optima.
b. It speeds up solving for θ using the normal equation.
c. It prevents the matrix $X^T X$ (used in the normal equation) from being non-invertable (singular/degenerate).
d. It speeds up gradient descent by making it require fewer iterations to get to a good solution.
Answer:
d
For a, the cost function $J(\theta)$ for linear regression has no local optima. For b, the magnitute of the feature values are significant in terms of computational cost. For c, feature scaling has nothing to do with matrix inversion. For d, feature scaling speeds up gradient descent by avoiding many extra iterations that are required when one or more features takes on much larger values than he rest.

10 Multiple choice II

In this part, some questions may have multiple correct answers.

1. Some of the problems below are best addressed using a supervised learning algorithm, and the others with an unsupervised learning algorithm. Which of the following would you apply supervised learning to? (Select all that apply.) In each case, assume some appropriate dataset is available for your algorithm to learn from.
   a. Given historical data of children's ages and heights, predict children's height as a function of their age.
   b. Given 50 articles written by male authors, and 50 articles written by female authors, learn to predict the gender of a new manuscript's author (when the identity of this author is unknown).
   c. Take a collection of 1000 essays written on the US Economy, and find a way to automatically group these essays into a small number of groups of essays that are somehow "similar" or "related".
   d. Examine a large collection of emails that are known to be spam email, to discover if there are sub-types of spam mail.

   Answer:
   ab

2. Some of the problems below are best addressed using a supervised learning algorithm, and the others with an unsupervised learning algorithm. Which of the following would you apply supervised learning to? (Select all that apply.) In each case, assume some appropriate dataset is available for your algorithm to learn from.
   a. Given data on how 1000 medical patients respond to an experimental drug (such as effectiveness of the treatment, side effects, etc.), discover whether there are different categories or "types" of patients in terms of how they respond to the drug, and if so what these categories are.
   b. Given a large dataset of medical records from patients suffering from heart disease, try to learn whether there might be different clusters of such patients for which we might tailor separate treatments.
   c. Have a computer examine an audio clip of a piece of music, and classify whether or not there are vocals (i.e., a human voice singing) in that audio clip, or if it is a clip of only musical instruments (and no vocals).
   d. Given genetic (DNA) data from a person, predict the odds of him/her developing diabetes over the next 10 years.

   Answer:
   cd

3. Some of the problems below are best addressed using a supervised learning algorithm, and the others with an unsupervised learning algorithm. Which of the following would you apply supervised learning to? (Select all that apply.) In each case, assume some appropriate dataset is available for your algorithm to learn from.
   a. Take a collection of 1000 essays written on the US Economy, and find a way to automatically group these essays into a small number of groups of essays that are somehow "similar" or "related".
   b. Given genetic (DNA) data from a person, predict the odds of him/her developing diabetes over the next 10 years.

c. Examine a large collection of emails that are known to be spam email, to discover if there are sub-types of spam mail.

d. Examine the statistics of two football teams, and predict which team will win tomorrow's match (given historical data of teams' wins/losses to learn from).

Answer:
bc

4. Which of these is a reasonable definition of machine learning?
   a. Machine learning is the science of programming computers.
   b. Machine learning learns from labeled data.
   c. Machine learning is the field of allowing robots to act intelligently.
   d. Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed.

Answer:
d

11 Extra Question

1. Which of the following statements about regularization are true? Check all that apply.
   a. Using a very large value of λ hurt the performance of your hypothesis; the only reason we do not set λ to be too large is to avoid numerical problems.
   b. Because logistic regression outputs values $0 \le h_\theta(x) \le 1$, its range of output values can only be "shrunk" slightly by regularization anyway, so regularization is generally not helpful for it.
   c. Consider a classification problem. Adding regularization may cause your classifier to incorrectly classify some training examples (which it had correctly classified when not using regularization, i.e. when $\lambda = 0$).
   d. Using too large a value of λ can cause your hypothesis to overfit the data; this can be avoided by reducing λ.
   Answer:
   c

2. Which of the following statements about regularization are true? Check all that apply.
   a. Using a very large value of λ hurt the performance of your hypothesis; the only reason we do not set λ to be too large is to avoid numerical problems.
   b. Because logistic regression outputs values $0 \le h_\theta(x) \le 1$, its range of output values can only be "shrunk" slightly by regularization anyway, so regularization is generally not helpful for it.
   c. Because regularization causes $J(\theta)$ to no longer be convex, gradient descent may
   d. Using too large a value of λ can cause your hypothesis to underfit the data; this can be avoided by reducing λ.
   Answer:
   d