CS777 – Week 3 Homework Submission Template

**!!!! PLEASE RENAME THIS DOCUMENT WITH YOUR NAME AND LASTNAME !!!!**

**Task 1 – Simple Linear Regression**

* Calculate the m slope and the b intercept for the small dataset

|  |
| --- |
| *Slope (m): 2.7014264077355357*  *Intercept (b): 3.9501934248399904* |

* Calculate the m slope and the b intercept for the large dataset

|  |
| --- |
| *Slope (m): 1.5276773830924568e-06*  *Intercept (b): 12.320230014656492*  *Time for the task: 1520.3017473220825 seconds* |

**Task 2 – Find the Parameters using Gradient Descent**

* Print out the costs and model parameters in each iteration. The maximum number of iterations is 50.

|  |
| --- |
| *Iteration 1, Cost: 42078078637.29253, m: 0.013392134708204325, b: 0.002464046719370065*  *Iteration 2, Cost: 41762264322.32871, m: 0.0266767912041712, b: 0.004919764461342796*  *Iteration 3, Cost: 41451403041.33091, m: 0.03985482535387381, b: 0.007367217780671859*  *Iteration 4, Cost: 41145416254.45652, m: 0.05292708620883528, b: 0.009806470718387597*  *Iteration 5, Cost: 40844226667.303444, m: 0.06589441606058564, b: 0.012237586805928638*  *Iteration 6, Cost: 40547758211.73824, m: 0.07875765049434796, b: 0.014660629069161995*  *Iteration 7, Cost: 40255936025.68076, m: 0.0915176184424087, b: 0.017075660032426245*  *Iteration 8, Cost: 39968686434.75286, m: 0.10417514223712927, b: 0.01948274172251482*  *Iteration 9, Cost: 39685936932.9688, m: 0.11673103766361663, b: 0.021881935672649133*  *Iteration 10, Cost: 39407616164.01667, m: 0.129186114011608, b: 0.02427330292638988*  *Iteration 11, Cost: 39133653903.7528, m: 0.14154117412743944, b: 0.02665690404155809*  *Iteration 12, Cost: 38863981041.67921, m: 0.153797014465101, b: 0.029032799094096923*  *Iteration 13, Cost: 38598529563.28537, m: 0.1659544251373076, b: 0.03140104768190485*  *Iteration 14, Cost: 38337232532.86926, m: 0.1780141899659497, b: 0.03376170892862277*  *Iteration 15, Cost: 38080024076.479485, m: 0.1899770865322021, b: 0.03611484148744676*  *Iteration 16, Cost: 37826839364.781715, m: 0.2018438862261172, b: 0.038460503544831784*  *Iteration 17, Cost: 37577614596.62993, m: 0.21361535429598422, b: 0.04079875282424831*  *Iteration 18, Cost: 37332286983.251686, m: 0.22529224989726715, b: 0.043129646589853216*  *Iteration 19, Cost: 37090794730.90396, m: 0.23687532614098236, b: 0.04545324165012498*  *Iteration 20, Cost: 36853077026.617485, m: 0.24836533014204656, b: 0.04776959436153438*  *Iteration 21, Cost: 36619074021.55717, m: 0.2597630030667971, b: 0.050078760632107866*  *Iteration 22, Cost: 36388726816.324265, m: 0.27106908018041664, b: 0.052380795925040954*  *Iteration 23, Cost: 36161977445.57443, m: 0.28228429089398355, b: 0.05467575526210871*  *Iteration 24, Cost: 35938768863.4719, m: 0.29340935881109065, b: 0.05696369322738827*  *Iteration 25, Cost: 35719044928.448456, m: 0.30444500177391925, b: 0.05924466397059212*  *Iteration 26, Cost: 35502750390.152084, m: 0.31539193190931736, b: 0.06151872121059078*  *Iteration 27, Cost: 35289830873.87992, m: 0.32625085567416484, b: 0.06378591823880116*  *Iteration 28, Cost: 35080232867.7309, m: 0.33702247390063267, b: 0.0660463079226198*  *Iteration 29, Cost: 34873903708.19741, m: 0.34770748184078787, b: 0.06829994270879014*  *Iteration 30, Cost: 34670791567.14981, m: 0.35830656921121473, b: 0.07054687462675266*  *Iteration 31, Cost: 34470845438.598495, m: 0.36882042023700884, b: 0.07278715529196186*  *Iteration 32, Cost: 34274015125.403133, m: 0.3792497136954547, b: 0.07502083590919185*  *Iteration 33, Cost: 34080251226.650314, m: 0.3895951229594532, b: 0.07724796727580976*  *Iteration 34, Cost: 33889505124.759754, m: 0.3998573160406011, b: 0.07946859978498819*  *Iteration 35, Cost: 33701728973.378033, m: 0.41003695563172493, b: 0.08168278342896065*  *Iteration 36, Cost: 33516875685.189587, m: 0.4201346991493891, b: 0.08389056780219376*  *Iteration 37, Cost: 33334898919.288074, m: 0.4301511987758357, b: 0.08609200210458183*  *Iteration 38, Cost: 33155753070.134583, m: 0.44008710150058755, b: 0.08828713514453516*  *Iteration 39, Cost: 32979393255.20645, m: 0.4499430491619716, b: 0.09047601534214132*  *Iteration 40, Cost: 32805775304.17099, m: 0.45971967848790024, b: 0.09265869073224811*  *Iteration 41, Cost: 32634855746.84085, m: 0.46941762113686025, b: 0.09483520896752294*  *Iteration 42, Cost: 32466591802.829693, m: 0.4790375037379708, b: 0.09700561732149937*  *Iteration 43, Cost: 32300941370.144966, m: 0.4885799479311279, b: 0.09916996269158541*  *Iteration 44, Cost: 32137863014.343945, m: 0.4980455704068439, b: 0.10132829160208495*  *Iteration 45, Cost: 31977315958.547085, m: 0.5074349829455349, b: 0.10348065020712763*  *Iteration 46, Cost: 31819260072.28109, m: 0.5167487924566746, b: 0.10562708429368133*  *Iteration 47, Cost: 31663655861.7508, m: 0.5259876010175308, b: 0.10776763928439505*  *Iteration 48, Cost: 31510464459.356373, m: 0.5351520059116356, b: 0.10990236024055522*  *Iteration 49, Cost: 31359647614.138195, m: 0.5442425996670294, b: 0.11203129186494623*  *Iteration 50, Cost: 31211167681.69621, m: 0.5532599700939819, b: 0.11415447850469411*  *Iteration 51, Cost: 31064987614.7232, m: 0.5622047003227744, b: 0.11627196415411757*  *Iteration 52, Cost: 30921070953.49367, m: 0.5710773688407021, b: 0.11838379245752657*  *Iteration 53, Cost: 30779381816.523823, m: 0.5798785495291974, b: 0.12049000671202234*  *Iteration 54, Cost: 30639884891.37964, m: 0.5886088117005452, b: 0.12259064987023158*  *Iteration 55, Cost: 30502545425.73514, m: 0.5972687201340955, b: 0.12468576454306933*  *Iteration 56, Cost: 30367329218.655888, m: 0.6058588351125133, b: 0.12677539300247018*  *Iteration 57, Cost: 30234202611.28706, m: 0.6143797124574342, b: 0.1288595771840692*  *Iteration 58, Cost: 30103132478.70002, m: 0.6228319035651807, b: 0.1309383586898868*  *Iteration 59, Cost: 29974086221.587482, m: 0.631215955441786, b: 0.13301177879097628*  *Iteration 60, Cost: 29847031757.31116, m: 0.6395324107380951, b: 0.13507987843007663*  *Iteration 61, Cost: 29721937512.404137, m: 0.647781807784372, b: 0.13714269822420158*  *Iteration 62, Cost: 29598772413.97894, m: 0.6559646806246884, b: 0.13920027846726443*  *Iteration 63, Cost: 29477505882.029762, m: 0.6640815590511179, b: 0.1412526591326141*  *Iteration 64, Cost: 29358107821.45435, m: 0.672132968637506, b: 0.14329987987562093*  *Iteration 65, Cost: 29240548614.651623, m: 0.6801194307732182, b: 0.14534198003619517*  *Iteration 66, Cost: 29124799113.654682, m: 0.6880414626961406, b: 0.14737899864128867*  *Iteration 67, Cost: 29010830632.72907, m: 0.6958995775260827, b: 0.1494109744073979*  *Iteration 68, Cost: 28898614941.235043, m: 0.7036942842973813, b: 0.151437945743031*  *Iteration 69, Cost: 28788124256.314877, m: 0.7114260879914077, b: 0.15345995075116758*  *Iteration 70, Cost: 28679331235.42021, m: 0.719095489568931, b: 0.1554770272316877*  *Iteration 71, Cost: 28572208970.16892, m: 0.7267029860021222, b: 0.15748921268377772*  *Iteration 72, Cost: 28466730978.50187, m: 0.7342490703062086, b: 0.15949654430832408*  *Iteration 73, Cost: 28362871198.604744, m: 0.741734231571088, b: 0.1614990590103149*  *Iteration 74, Cost: 28260603982.00768, m: 0.7491589549925681, b: 0.16349679340113688*  *Iteration 75, Cost: 28159904086.914314, m: 0.7565237219032557, b: 0.16548978380100018*  *Iteration 76, Cost: 28060746671.954437, m: 0.7638290098034692, b: 0.16747806624117162*  *Iteration 77, Cost: 27963107289.806374, m: 0.771075292391676, b: 0.16946167646633215*  *Iteration 78, Cost: 27866961880.782806, m: 0.7782630395948019, b: 0.17144064993681585*  *Iteration 79, Cost: 27772286766.78723, m: 0.7853927175981826, b: 0.17341502183091353*  *Iteration 80, Cost: 27679058645.312252, m: 0.7924647888753938, b: 0.1753848270470915*  *The function took 1580.9004516601562 seconds* |

* Calculate the m slope and the b intercept based on your calculations

|  |
| --- |
| *m: 0.7924647888753938, b: 0.1753848270470915* |

* Comment on how you can interpret the parameters of the model. What is the meaning of m and b in this case?

|  |
| --- |
| *for each unit increase in the independent variable x, the dependent variable y is expected to increase by 0.7925 units.*  *The value of* b*=0.1754 is the expected value of* y *when* x *is zero.* |

**Task 3 – Fit Multiple Linear Regression using Gradient Descent**

* Print out the costs and model parameters in each iteration.

|  |
| --- |
| *It converges at 53* |

* What are m and b values based on your calculations?

|  |
| --- |
| *m: [ 1.22136650e+19 -5.47310795e+22 -1.80969516e+23 -5.66001423e+21],*  *b: -9.170675401577221e+21,* |

* Comment on how you can interpret the parameters of the model. What is the meaning of *mi* and *b* in this case

|  |
| --- |
| *for each unit increase in the independent variable x1,x2,x3,x4(holding all other variables constant), the dependent variable y is expected to increase by 1.22136650e+19, -5.47310795e+22,*  *-1.80969516e+23, -5.66001423e+21units.*  *The value of* b*=-9.170675401577221e+21 is the expected value of* y *when* x *is zero.* |

**Spark History Output:**

**Task 1:**

|  |
| --- |
|  |

**Task 2:**

|  |
| --- |
|  |

**Task 3:**

|  |
| --- |
| *I run task2 3 together, but it is too long to capture all, it may due to the additional clean steps that consume one more hour to run. Since if I do not add some additional steps to clean the data, the cost in task2 will never converge but diverge instead. However, even after adding clean steps, it can still not converge after 80 iterations which made me confused a lot. I tried many times and spent all night on it but still cannot find a good way to solve this. It is real time consuming even I still have some thoughts but do not have enough time to try.* |