Midterm preparation

The dataset about diamonds is shown on the figure below.

If we want to predict price based on other diamond parameters, we are talking about supervised learning regression problem, where our target feature (label) is price. There are 10 features: unnamed (numerical, discrete), carat (numerical, continuous), cut (categorical), color (categorical), clarity (categorical), depth (numerical, continuous), table (numerical continuous), x (numerical continuous), y (numerical continuous), and z (numerical continuous). There are 53940 observations. We can model this problem using multivariate linear regression with gradient descent optimization. Hypothesis for MLR is:

$h_\theta(X) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$

where θ_0, θ_1, θ_2, ..., θ_n are weights and x_1, x_2, ..., x_n are features (all non-numerical features should be converted to numerical).

Cost function: $J(\theta_0, \theta_1, \theta_2, \dots, \theta_n) = \frac{1}{2m} \sum_{i=0}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$

Gradient descent algorithm:

Repeat until convergence {

$$\theta_j := \theta_j - \alpha \frac{2}{m} \sum_{i=0}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

(simultaneously update θ_j for $j = 0, 1, 2, \dots, n$)}

	Unnamed: 0	carat	cut	color	clarity	depth	table	price	x	y	z
0	1	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	2	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	3	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	4	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	5	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
...	...	...	...	...	...	...	...	...	...	...	...
53935	53936	0.72	Ideal	D	SI1	60.8	57.0	2757	5.75	5.76	3.50
53936	53937	0.72	Good	D	SI1	63.1	55.0	2757	5.69	5.75	3.61
53937	53938	0.70	Very Good	D	SI1	62.8	60.0	2757	5.66	5.68	3.56
53938	53939	0.86	Premium	H	SI2	61.0	58.0	2757	6.15	6.12	3.74
53939	53940	0.75	Ideal	D	SI2	62.2	55.0	2757	5.83	5.87	3.64

53940 rows × 11 columns

If we are predicting the diamond color, we are talking about a supervised learning multinomial classification problem, where our target feature is color with labels D, E, F, G, H, I, and J. To model this problem we can use logistic regression algorithm. In case we analyze this as one vs all classification problem (i.e. we are predicting if the color is either D or any else), the hypothesis can be formulated as

$$h_\theta(x) = \frac{1}{1-e^{-\theta^T x}}$$

Weights (parameters): $\theta_0, \theta_1, \theta_2, ..., \theta_n$

Features: $x_1, x_2, ..., x_n$

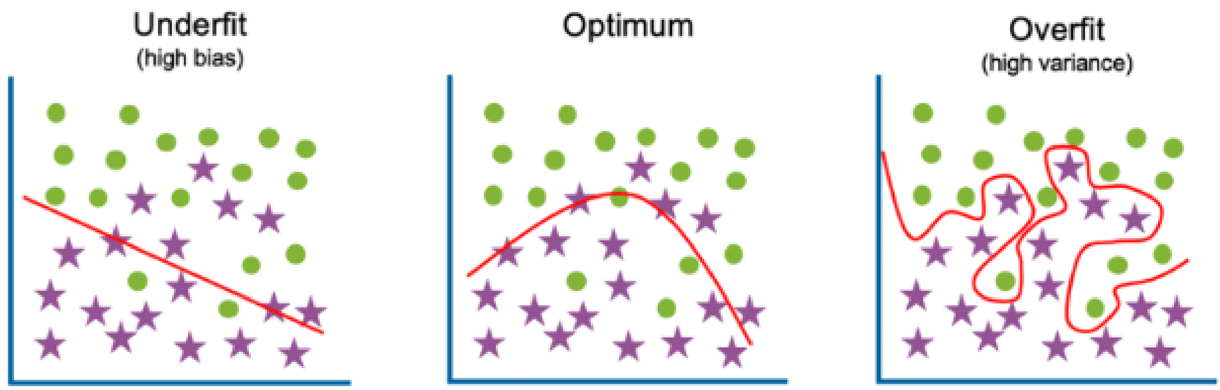
Cost function: $J(\theta) = -\frac{1}{m}[\sum_{i=0}^{m} y^{(i)} log\, h_\theta(x^{(i)}) + (1 - y^{(i)}) log\ (1 - log\, h_\theta(x^{(i)}))]$

Gradient descent algorithm:

Repeat until convergence {

$$\theta_j := \theta_j - \alpha \frac{2}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

(simultaneously update θ_j for $j = 0, 1, 2, ..., n$)}



To prevent underfitting, decrease the value of the regularization parameter, increase the number of features, and add polynomial features.

To prevent overfitting, increase the value of the regularization parameter, decrease the number of features, and increase the number of observations.