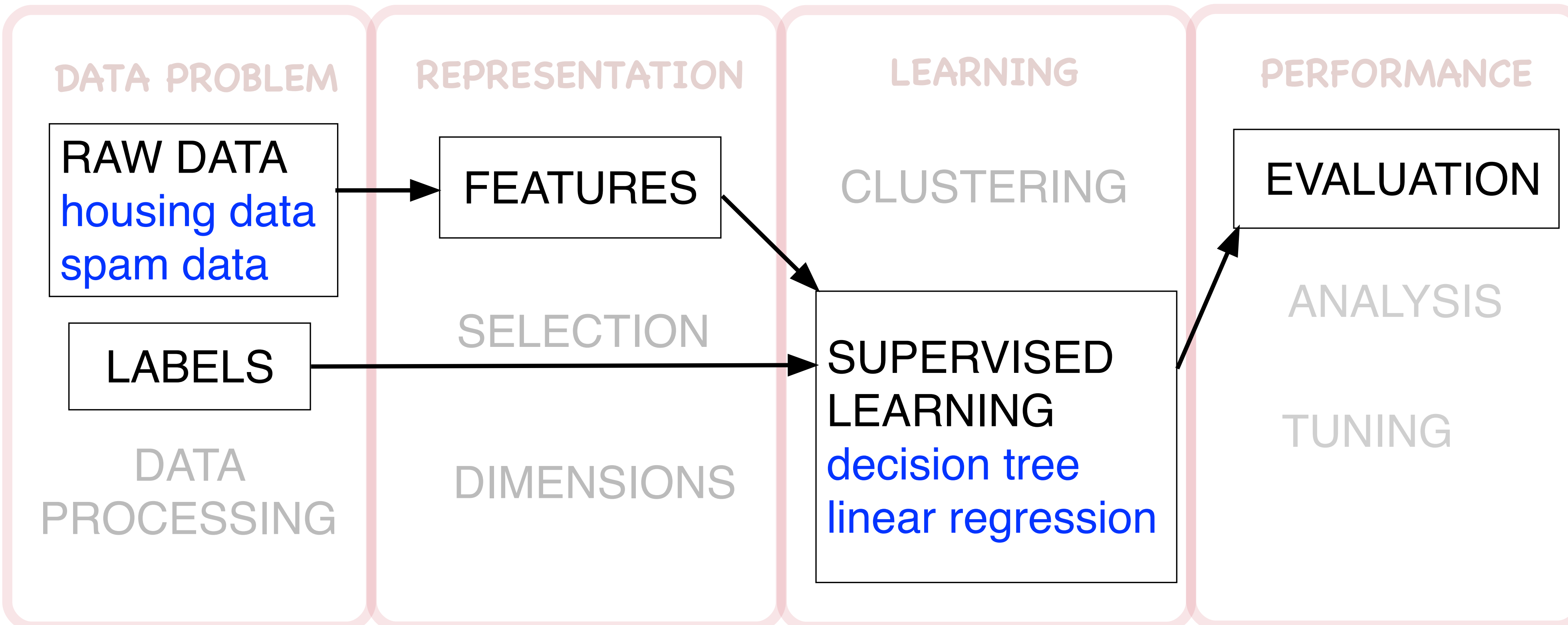
Linear Regression via Normal Equations

some material thanks to Andrew Ng @Stanford

Course Map / module1

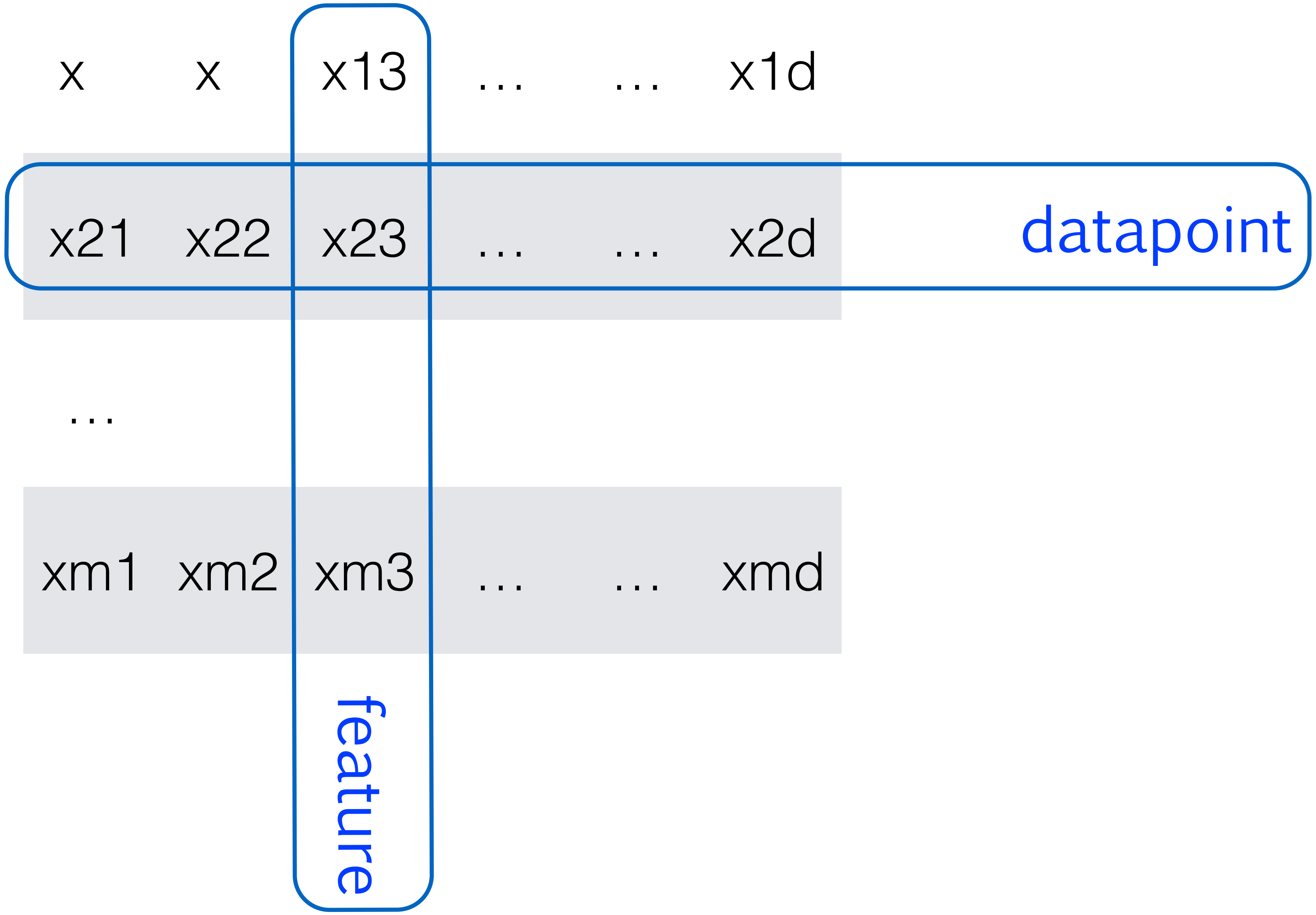


- two basic supervised learning algorithms
  - decision trees
  - linear regression
- two simple datasets
  - housing
  - spam emails

Module 1 Objectives / Linear Regression

- Linear Algebra Primer
  - matrix equations, notations
  - matrix manipulations
- Linear Regression
  - objective, convexity
  - matrix form
  - derivation of normal equations
- Run regression in practice

Matrix data



- m datapoints/objects Xi=(x1,x2,…,xd); i=1:m
- d features/columns f1, f2, …, fd
- label(X_i) = y_i, given for each datapoint in the training set.

Matrix data / training VS testing

Training

	AUT	BEL	BUL	CYP	CZE	DEN	EST	FIN	FRA	GER	GRE	HUN	IRL	ITA	LAT	LTU	LUX	MLT	NED	POL	POR	ROM	SVK	SLO	ESP	SWE	GBR
T-01	64.4	125.0	44.7	7.0	124.1	51.3	14.9	56.6	363.5	837.4	92.2	56.8	42.8	446.6	6.5	11.6	8.4	2.1	174.8	303.8	64.8	90.7	36.9	15.1	304.9	48.8	558.2
T-02	7.1	7.8	10.3	0.7	11.0	5.6	1.9	4.5	56.9	47.6	9.3	7.8	13.1	39.8	1.8	3.3	0.3	0.3	16.7	38.3	11.4	25.7	4.2	2.1	37.3	5.6	49.5
T-03	5.3	11.0	4.4	0.7	8.0	7.0	0.8	6.9	72.3	66.5	9.1	9.7	8.8	40.5	1.5	5.0	0.4	0.3	17.6	31.1	6.1	16.8	3.7	1.3	29.6	7.7	39.6
T-04	118	141	90	10	10	14	16	10	1,801	718	128	209	174	361	3	41	6	5	265	261	129	570	20	124	244	296	351
T-05	912	1,454	387	91	594	805	8	864	10,958	9,363	1,162	518	431	5,267	19	19	83	42	1,354	2,750	391	4	175	95	5,011	777	9,221
T-06	287	43	4	16	86	22	6	20	1,354	4,740	210	201	96	460	8	1	4	8	337	24	10	0	17	19	272	142	1,143
T-07	644	1,250	447	70	1,241	513	149	566	3,635	8,374	922	568	428	4,466	65	116	84	21	1,748	3,038	648	907	369	151	3,049	488	5,582
T-08	782	1,126	480	82	779	988	120	558	9,533	6,354	1,045	846	1,845	3,721	192	405	38	38	1,817	3,488	824	2,028	322	202	4,476	857	4,489
T-09	228	133	648	26	291	137	53	244	1,410	1,369	328	394	178	1,933	76	154	3	12	664	1,221	647	740	211	65	1,296	215	2,208
T-10	832	1,046	764	86	1,033	546	134	530	6,410	8,231	1,115	1,005	430	5,921	23	337	47	41	1,639	3,813	1,062	2,144	539	202	4,512	915	6,059
T-11	305	11	112	8	125	109	89	297	619	1,166	43	83	16	338	97	59	4	1	95	732	47	58	110	15	466	319	255
T-12	501	467	314	448	373	354	350	448	491	546	348	280	385	581	297	384	659	525	429	314	572	149	222	456	454	456	463
T-13	282	641	131	53	203	171	60	220	1,970	2,650	436	132	182	1,881	47	56	62	19	947	446	332	212	74	53	1,573	362	1,827
T-14	65.2	82.4	37.4	4.5	58.8	36.4	6.8	80.8	482.4	524.6	53.5	37.1	23.2	303.8	6.3	9.4	6.1	2.1	102.4	124.1	46.1	49.6	28.6	13.7	241.8	137.8	345.2
T-15	9.00	17.06	3.47	0.01	9.60	4.82	1.44	4.86	45.41	102.00	2.34	14.46	4.30	80.61	1.91	2.92	1.36	0.00	51.30	15.67	4.30	18.00	6.00	1.10	27.01	0.98	98.47
T-16	3.00	3.10	7.40	0.00	19.40	5.50	0.00	5.20	13.10	82.40	8.80	2.90	0.00	17.40	0.00	0.20	3.10	0.00	7.50	58.40	3.70	7.60	3.80	0.00	18.30	2.20	43.80
T-17	369	989	98	89	389	385	8	77	10,979	3,463	999	770	233	16,980	53	60	55	30	1,492	950	1,246	270	280	950	3,402	179	3,313
T-18	227	289	157	23	395	317	42	297	4,178	2,612	420	323	573	1,681	64	162	0	1	409	1,557	228	327	120	72	2,183	287	1,909
T-19	3.5	5.8	2.3	3.2	3.9	3.6	3.3	6.4	4.4	4.1	2.6	2.4	3.9	3.0	1.5	2.0	8.4	2.1	4.8	2.3	2.5	1.6	3.2	3.3	3.1	5.4	4.0
T-20	6.9	7.7	3.3	5.3	5.4	6.2	4.5	15.5	6.8	6.3	4.6	3.2	5.9	2.3	1.7	1.3	13.5	1.3	6.4	1.5	1.8	1.1	2.0	2.4	2.3	3.7	2.5
T-21	0.46	3.43	0.19	0.00	0.43	1.01	0.09	0.19	0.99	1.82	0.47	0.23	0.45	1.00	0.04	0.21	0.00	0.00	1.51	0.28	0.22	0.27	0.27	0.17	0.44	0.27	2.76

Testing

	AUT	BEL	BUL	CYP	CZE	DEN	EST	FIN	FRA	GER	GRE	HUN	IRL	ITA	LAT	LTU	LUX	MLT	NED	POL	POR	ROM	SVK	SLO	ESP	SWE	GBR
T-22	29	38	48	100	76	83	100	39	8	62	95	60	96	79	29	17	57	100	90	98	65	63	30	35	50	4	74
T-23	133	178	7	13	44	111	8	129	786	782	103	32	164	395	11	10	38	15	227	72	96	20	2	13	518	234	985
T-24	804	334	65	192	471	1,034	58	708	5,248	9,079	945	274	4,287	3,612	103	51	85	137	2,613	355	1,014	171	71	76	4,986	902	9,360
T-25	130	103	7	0.00	53	78	7	97	860	1,070	80	46	197	398	7	10	74	22	429	68	128	26	5	13	473	129	977
T-26	0.13	0.19	0.10	0.12	0.57	0.12	0.05	0.10	0.32	0.31	0.10	0.22	0.17	0.36	0.13	0.14	0.19	0.10	0.28	0.35	0.17	0.10	0.27	0.29	0.15	0.17	0.27
T-27	630	464	463	739	289	737	436	468	543	601	438	459	740	542	310	378	705	611	624	245	446	382	289	423	597	482	584
T-28	46	17	4	5	4	8	0	47	59	17	6	4	27	47	0	1	0	0	19	31	15	7	26	16	85	31	62
T-29	521	828	1,004	3,711	1,359	843	1,254	697	162	1,140	2,247	976	2,423	1,473	362	139	1,707	2,856	1,575	1,501	1,377	744	798	851	1,248	41	1,170
T-30	347	330	107	230	220	371	203	335	312	319	240	175	445	302	160	153	714	213	321	144	198	91	186	234	274	322	318
T-31	0.0	0.0	20.2	4.8	0.6	7.0	0.6	0.2	20.5	0.1	18.3	1.3	0.5	76.4	1.0	0.1	0.1	0.1	0.2	49.5	111.8	1.6	0.1	0.7	92.4	1.3	0.2
T-32	24.7	20.1	34.1	13.3	34.3	21.6	24.1	28.8	16.4	26.3	0.0	29.6	26.7	22.6	18.3	30.1	9.7	20.2	20.8	29.5	20.9	36.1	32.5	29.7	21.1	25.4	18.4
T-33	134	117	34	8	127	72	13	51	951	231	107	70	96	480	28	69	5	2	105	249	59	98	37	20	657	167	372
T-34	9.2	37.0	6.3	0.0	8.1	7.5	0.0	12.8	86.3	122.7	21.2	8.4	3.1	100.6	0.0	9.2	0.0	0.0	84.7	18.5	13.6	14.9	6.2	0.0	60.3	19.8	86.0
T-35	1.0	5.2	0.5	0.1	1.4	0.3	7.3	2.5	7.6	20.0	0.3	1.4	0.7	6.1	0.0	0.1	0.1	0.0	1.9	1.6	2.3	2.2	0.4	0.1	3.1	1.2	8.0

regression goal

- housing data, two features (toy example)

Living area (ft^2)	#bedrooms	price (1000$s)
2104	3	400
1600	3	330
2400	3	369
1416	2	232
3000	4	540
...	...	...

- regressor = a linear predictor

$$h_\theta(\mathbf{x}) = \theta^0 + \theta^1 x^1 + \theta^2 x^2$$

$$h(\mathbf{x}) = \sum_{d=0}^{D} \theta^d x^d$$

- such that h(x) approximates label(x)=y as close as possible, measured by square error

$$J(\theta) = \sum_t (h_\theta(\mathbf{x}_t) - y_t)^2$$

Regression Normal Equations

- Linear regression has a well known exact solution, given by linear algebra

- X= training matrix of feature values
- Y= corresponding labels vector

- then regression coefficients that minimize objective J are

$$\theta = (X^T X)^{-1} X^T Y$$

Normal equations: matrix derivatives

- if function f takes a matrix and outputs a real number, then its derivative is

$$\nabla_A f(A) = \begin{bmatrix} \frac{\partial f}{\partial a_{11}} & \cdots & \frac{\partial f}{\partial a_{1n}} \\ \cdots & \cdots & \cdots \\ \frac{\partial f}{\partial a_{m1}} & \cdots & \frac{\partial f}{\partial a_{mn}} \end{bmatrix}$$

- example:

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

$$f(A) = \frac{3}{2}a_{11} + 5a_{12}^2 + a_{21}a_{22} \qquad \nabla_A f(A) = \begin{bmatrix} \frac{3}{2} & 10a_{12} \\ a_{22} & a_{21} \end{bmatrix}$$

Normal equations : matrix trace

- trace(A) = sum of main diagonal

$$\mathrm{tr}(A) = \sum_i a_{ii}$$

- easy properties

$$\begin{aligned} \mathrm{tr}(AB) &= \mathrm{tr}(BA) \\ \mathrm{tr}(A) &= \mathrm{tr}(A^T) \\ \mathrm{tr}(A+B) &= \mathrm{tr}(A) + \mathrm{tr}(B) \\ \mathrm{tr}(xA) &= x\mathrm{tr}(A) \end{aligned}$$

- advanced properties

$$\begin{aligned} \nabla_A \mathrm{tr}(AB) &= B^T \\ \nabla_{A^T} f(A) &= (\nabla_A f(A))^T \\ \nabla_A \mathrm{tr}(ABA^T C) &= CAB + C^T AB^T \end{aligned}$$

$$\nabla_{A^T} \mathrm{tr}(ABA^T C) = B^T A^T C^T + BA^T C$$

regression checkpoint: matrix derivative and trace

- 1) in the example few slides ago explain how the matrix of derivatives was calculated

$$f(A) = \frac{3}{2}a_{11} + 5a_{12}^2 + a_{21}a_{22}$$

$$\nabla_A f(A) = \begin{bmatrix} \frac{3}{2} & 10a_{12} \\ a_{22} & a_{21} \end{bmatrix}$$

- 2) derive on paper the first three advanced matrix trace properties

$$\nabla_A \mathrm{tr}(AB) = B^T$$

$$\nabla_{A^T} f(A) = (\nabla_A f(A))^T$$

$$\nabla_A \mathrm{tr}(ABA^T C) = CAB + C^T AB^T$$

Normal equations : mean square error

- data and labels

$$X = \begin{bmatrix} x_1^1 & \dots & x_1^D \\ \dots & \dots & \dots \\ x_m^1 & \dots & x_m^D \end{bmatrix} \quad Y = \begin{bmatrix} y_1 \\ \dots \\ y_m \end{bmatrix}$$

- error (difference) for regressor

$$E = \begin{bmatrix} h_\theta(\mathbf{x}_1) - y_1 \\ \dots \\ h_\theta(\mathbf{x}_m) - y_m \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1\theta \\ \dots \\ \mathbf{x}_m\theta \end{bmatrix} - \begin{bmatrix} y_1 \\ \dots \\ y_m \end{bmatrix} = X\theta - Y$$

- square error

$$J(\theta) = \frac{1}{2}\sum_t (h_\theta(\mathbf{x}_t) - y_t)^2 = \frac{1}{2}E^T E = \frac{1}{2}(X\theta - Y)^T(X\theta - Y)$$

Normal equations : mean square error differential

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \frac{1}{2} E^T E = \nabla_\theta \frac{1}{2} (X\theta - Y)^T (X\theta - Y) \\
&= \frac{1}{2} \nabla_\theta (\theta^T X^T X\theta - \theta^T X^T Y - Y^T X\theta + Y^T Y) \\
&= \frac{1}{2} \nabla_\theta \mathrm{tr}(\theta^T X^T X\theta - \theta^T X^T Y - Y^T X\theta + Y^T Y) \\
&= \frac{1}{2} \nabla_\theta (\mathrm{tr}(\theta^T X^T X\theta) - 2\mathrm{tr}(Y^T X\theta)) \\
&= \frac{1}{2} (X^T X\theta + X^T X\theta - 2X^T Y) \\
&= X^T X\theta - X^T Y
\end{aligned}
$$

- minimize J =>set the derivative to zero:

$$X^T X\theta = X^T Y \text{ or } \theta = (X^T X)^{-1} X^T Y$$

linear regression: use on test points

- $x=(x^1,x^2,\ldots,x^d)$ test point
- $h = (\theta^0,\theta^1,\ldots,\theta^d)$ regression model
- apply regressor to get a predicted label (add bias feature $x^0=1$)

$$h(\mathbf{x}) = \sum_{d=0}^{D} \theta^d x^d$$

- if y=label(x) is given, measure error
  - absolute difference $|y-h(x)|$
  - square error $(y-h(x))^2$

Logistic regression

- Logistic transformation

$$g(z) = \frac{1}{1+e^{-z}}$$

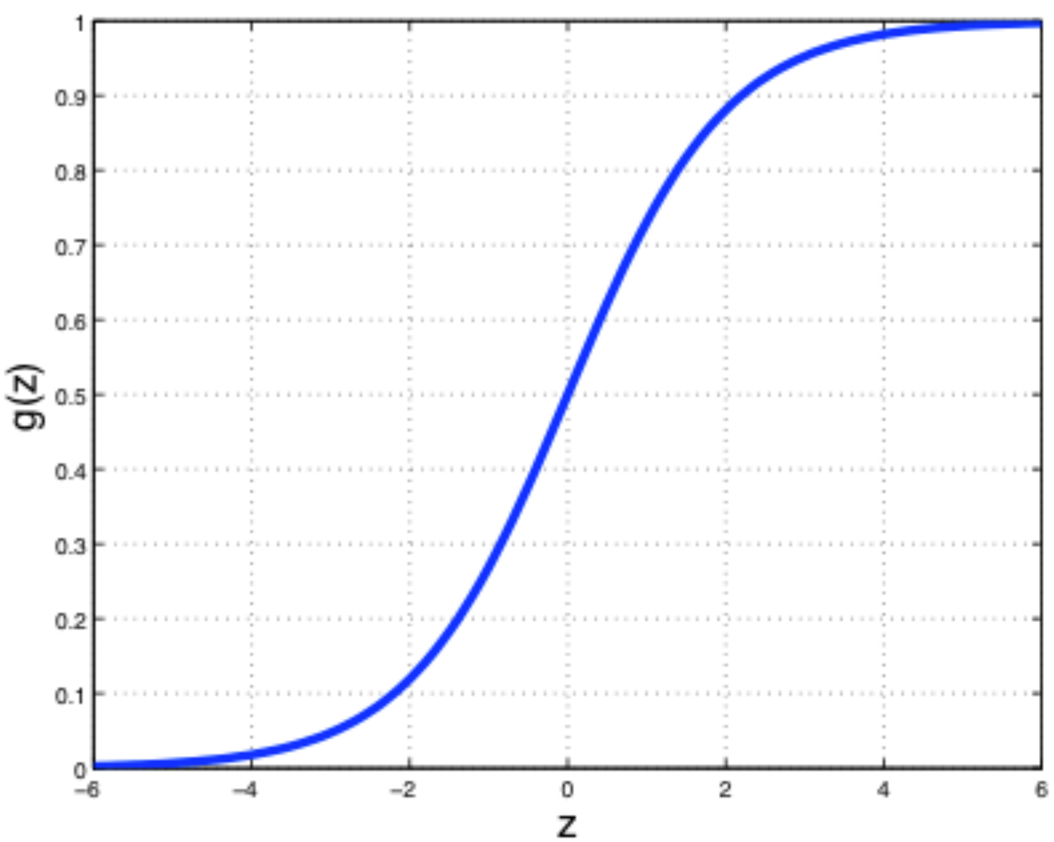


Figure 1: Logistic function

- Logistic differential

$$\begin{aligned} g'(z) &= \frac{\partial g(z)}{\partial z} \\ &= \frac{1}{(1+e^{-z})^2} e^{-z} \\ &= \frac{1}{1+e^{-z}} \left(1 - \frac{1}{1+e^{-z}}\right) \\ &= g(z)(1-g(z)) \end{aligned}$$

Logistic regression

- Logistic regression function

$$h_w(\mathbf{x}) = g(w\mathbf{x}) = \frac{1}{1 + e^{-w\mathbf{x}}} = \frac{1}{1 + e^{-\sum_d w^d x^d}}$$

- Solve the same optimization problem as before
  - no exact solution this time, will use gradient descent (numerical methods) next module

Linear Regression Screencast

- http://www.screencast.com/t/U3usp6TyrOL