Review Questions

Verena Blaschke

June 06, 2018

10

The Kullback-Leibler divergence of two distributions P and Q, $D_{KL}(P \parallel Q)$, is always larger than the entropy of P, $H(P)$.

10

The Kullback-Leibler divergence of two distributions P and Q, $D_{KL}(P \parallel Q)$, is always larger than the entropy of P, $H(P)$.

False

10

The Kullback-Leibler divergence of two distributions P and Q, $D_{KL}(P \parallel Q)$, is always larger than the entropy of P, $H(P)$.

False

$$D_{KL}(P \parallel Q) = H(P, Q) - H(P)$$

Entropy $H(P)$:
How costly is it to encode data from this distribution?

Cross entropy $H(P, Q)$:
entropy of (the true distribution) P under (its approximation) Q

KL divergence $D_{KL}(P \parallel Q)$:
How much more costly is it to encode data from P using Q?

10

The Kullback-Leibler divergence of two distributions P and Q, $D_{KL}(P \parallel Q)$, is always larger than the entropy of P, $H(P)$.

False

$$D_{KL}(P \parallel Q) = H(P, Q) - H(P)$$

$$H(P) = -\sum_x P(x) log P(x)$$

$$H(P, Q) = -\sum_x P(x) log Q(x)$$

Example: $P(x = A) = 0.5$ $\quad$ $Q(x = A) = 0.4$

$P(x = B) = 0.5$ $\quad$ $Q(x = B) = 0.6$

$$H(P) = -(0.5 \times log_2 0.5 + 0.5 \times log_2 0.5) = 1$$

$$H(P, Q) = -(0.5 \times log_2 0.4 + 0.5 \times log_2 0.6)$$
$$= -(0.5 * -1.32 + 0.5 * -0.74) = 1.03$$

$$D_{KL}(P \parallel Q) = 1.03 - 1 = 0.03$$

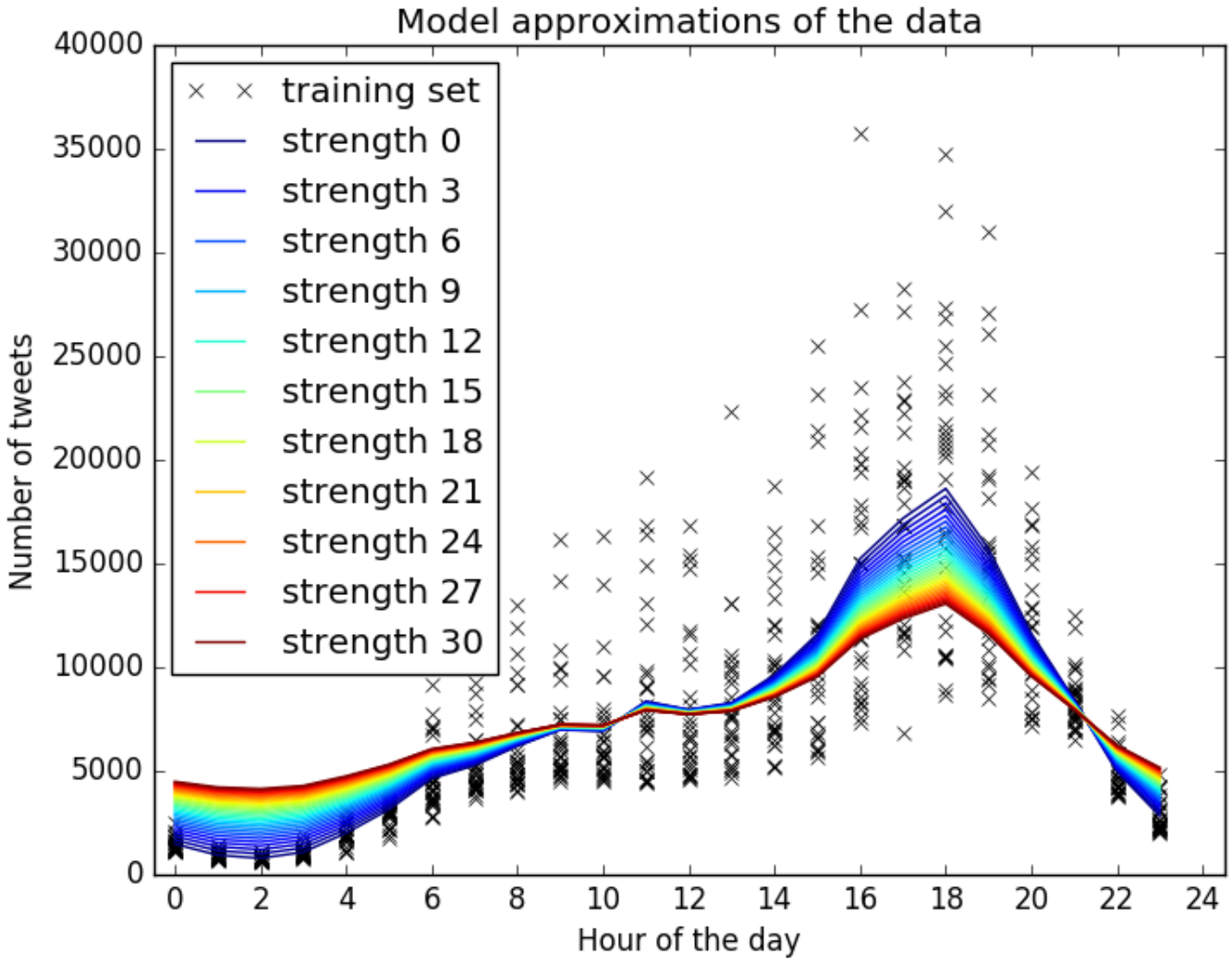
11

A regularized estimation of a machine learning model reduces the model's fit to the training data.

11

A regularized estimation of a machine learning model reduces the model's fit to the training data.

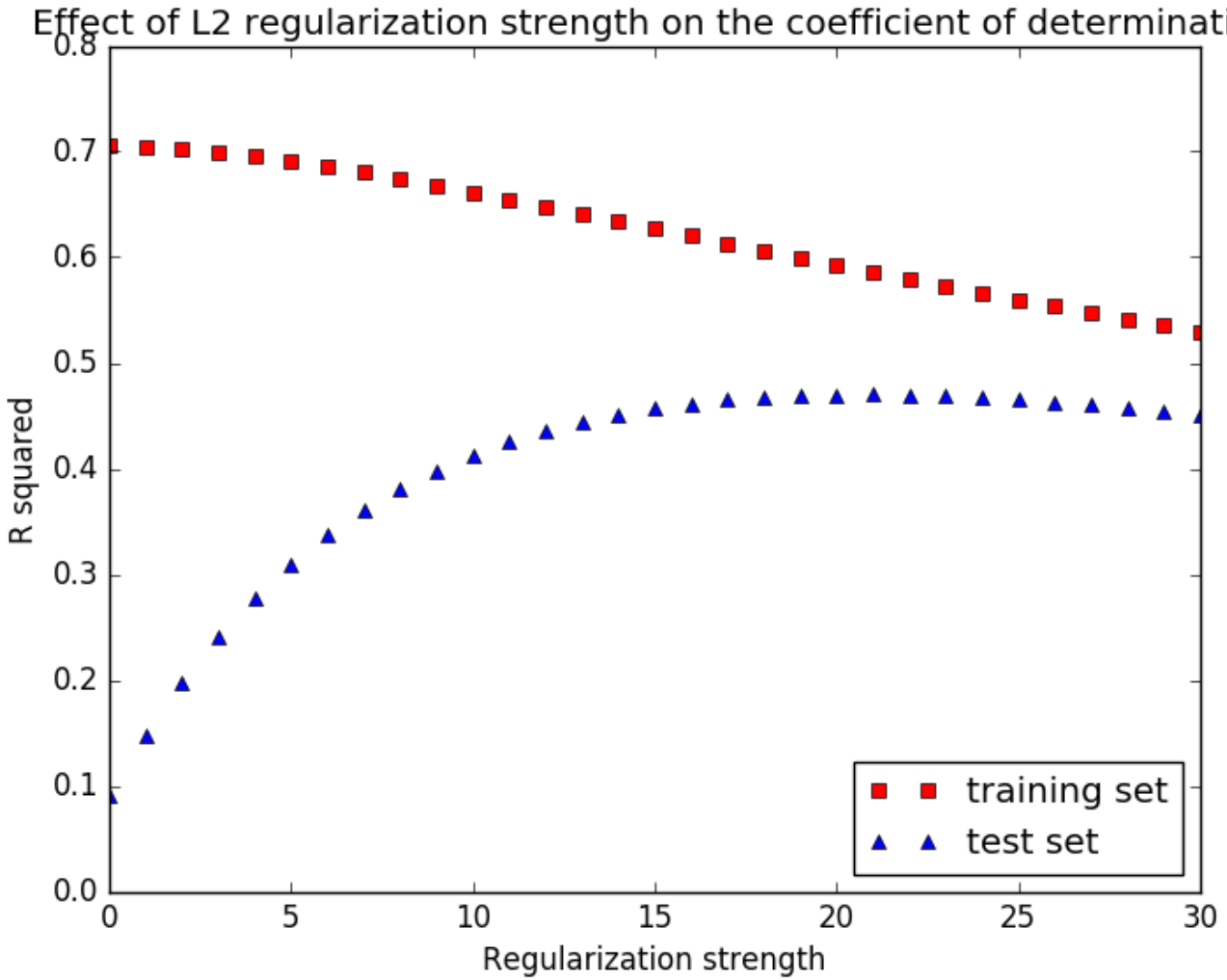
True



11

A regularized estimation of a machine learning model reduces the model's fit to the training data.

True



12

If a machine learning model has a convex loss function, one can calculate the minimum loss analytically.

12

If a machine learning model has a convex loss function, one can calculate the minimum loss analytically.

False

- analytic solutions are possible for **some** convex loss functions (e.g. least squares regression)
- ...but not all

12

If a machine learning model has a convex loss function, one can calculate the minimum loss analytically.

False

- analytic solutions are possible for **some** convex loss functions (e.g. least squares regression)
- ...but not all
- but search procedures (e.g. gradient descent) can find the minimum

13

No correlation with the outcome variable is a desired property of the predictors for a statistical model.

13

No correlation with the outcome variable is a desired property of the predictors for a statistical model.

False

Correlation with the outcome variable is *exactly* what we want!

14

The perceptron algorithm adjusts the weights after every correctly classified training sample.

14

The perceptron algorithm adjusts the weights after every correctly classified training sample.

False

While training the model, if an instance is...

- classified correctly, nothing happens.
- misclassified, the weights are updated.

15

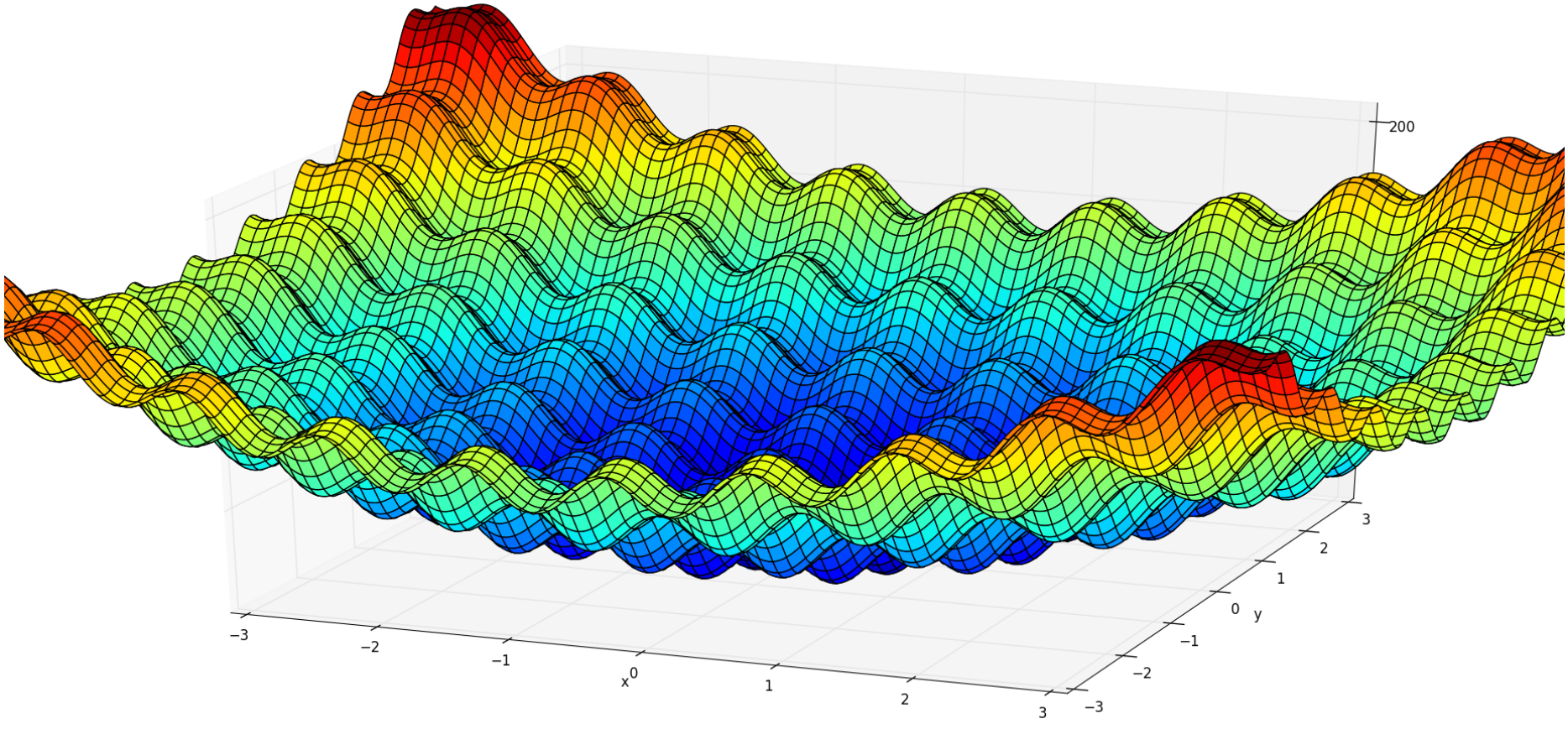
The gradient of a multivariate function is the 0 vector only at the global minimum of the function.

15

The gradient of a multivariate function is the 0 vector only at the global minimum of the function.

False

The gradient is 0 at any global or local minimum, of which there can be many:



16

Multiplying a matrix with its eigenvector does not change the direction of the vector.

16

Multiplying a matrix with its eigenvector does not change the direction of the vector.

True

matrix $\boldsymbol{A}$ with an eigenvector-eigenvalue pair v and λ:

$$\boldsymbol{A}v = \lambda v$$

same direction, different size

17

A machine learning system with high recall is likely to produce few false positives.

17

A machine learning system with high recall is likely to produce few false positives.

False

prediction \ *true value*	pos.	neg.
pos.	TP	FP
neg.	FN	TN

$$recall = \frac{TP}{TP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

17

A machine learning system with high recall is likely to produce few false positives.

False

		true value	
		pos.	neg.
prediction	pos.	TP	FP
	neg.	FN	TN

$$recall = \frac{TP}{TP + FN} \qquad precision = \frac{TP}{TP + FP}$$

Example: 1000 documents, of which 3 are relevant

If a system returns all documents: $FP = 9997$

$$recall = \frac{3}{3 + 0} = 100\% \qquad precision = \frac{3}{3 + 9997} = 0.3\%$$

18

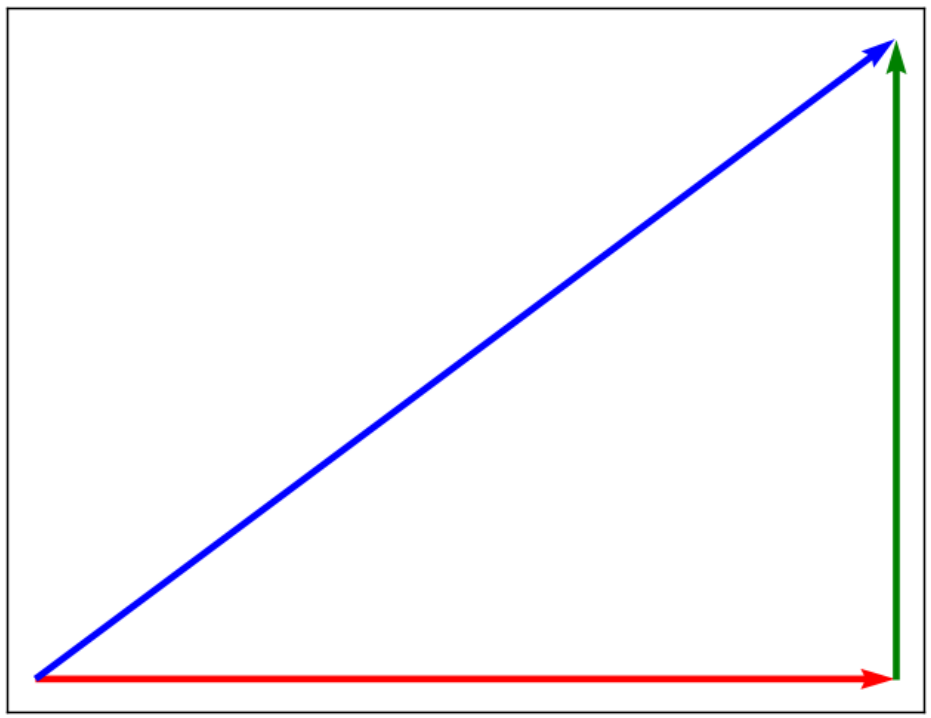
The L_2 norm of a vector in R^n (for any n in range $(1, \infty)$) is always smaller than or equal to its L_1 norm.

18

The L_2 norm of a vector in R^n (for any n in range $(1, \infty)$) is always smaller than or equal to its L_1 norm.

True

$$\|v\|_1 = \sum_{i=1}^{n} |v_i| \qquad \|v\|_2 = \sqrt{\sum_{i=1}^{n} |v_i|^2}$$



19

If the mutual information between two random variables x and y is $MI(X, Y) = 0$, the conditional entropy is $H(Y|X) = H(Y)$.

19

If the mutual information between two random variables x and y is $MI(X, Y) = 0$, the conditional entropy is $H(Y|X) = H(Y)$.

True

$MI(X, Y) = 0$ when X and Y are independent
$(P(X, Y) = P(X)P(Y))$.

When X and Y are independent, knowing about X doesn't give us any information about X:
$H(Y|X) = H(Y)$ (and vice versa).

20

L_1 regularization results in sparse parameter estimates.

20

L_1 regularization results in sparse parameter estimates.

True

Minimizing

$$J(w) + \lambda \sum_{j=1}^{k} |w_j|$$

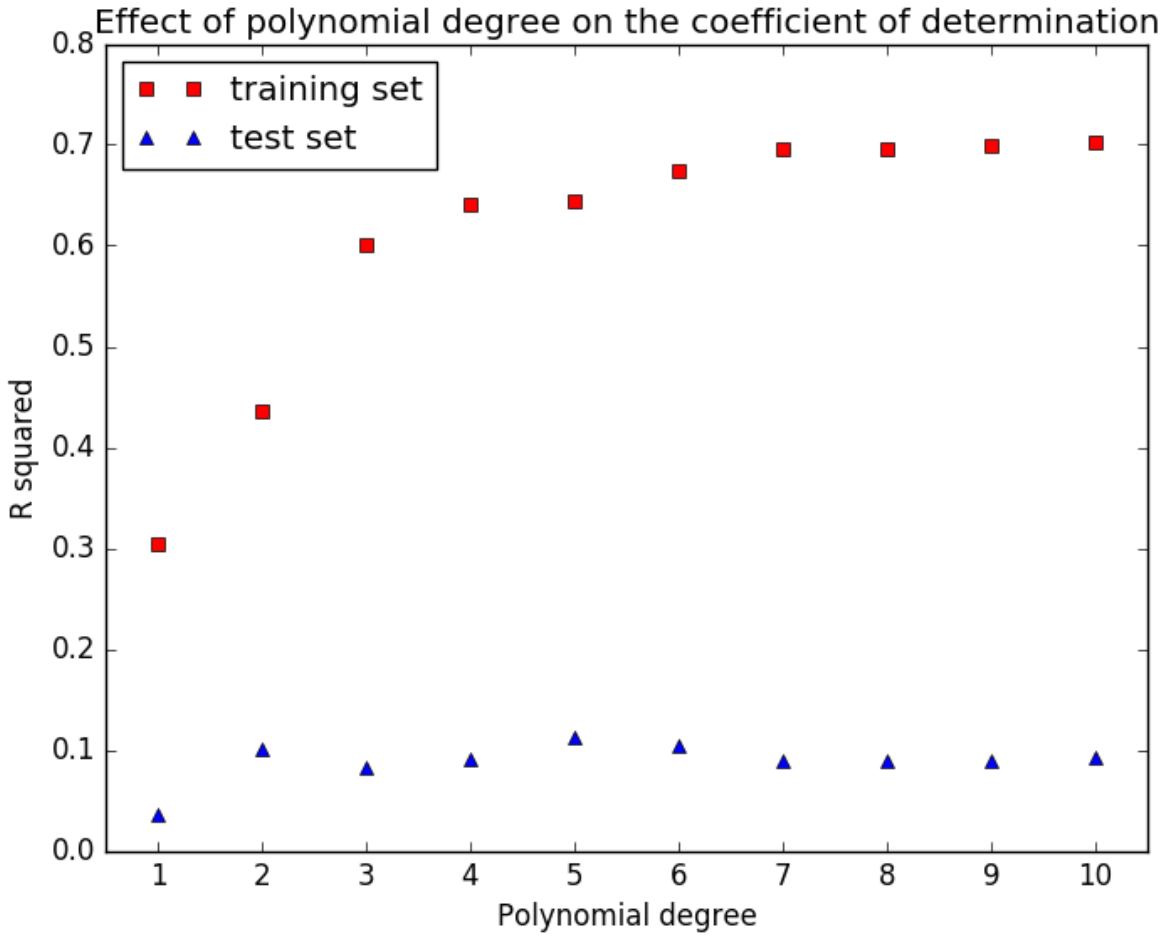
tends to result in a model where the less important features' coefficients are 0.

Increasing model complexity (e.g., number of parameters) in a machine learning model is likely to decrease test error.

Increasing model complexity (e.g., number of parameters) in a machine learning model is likely to decrease test error.

False

Overfitting to training data



Appendix

Entropy, Cross Entropy, KL Divergence

Entropy: How costly is it to encode data from this distribution?

$$\begin{aligned} H(P) &= \sum_x P(x) log \frac{1}{P(x)} \\ &= -\sum_x P(x) log P(x) \end{aligned} \tag{1}$$

Cross entropy:
entropy of (the true distribution) P under (its approximation) Q

$$H(P, Q) = -\sum_x P(x) log Q(x) \tag{2}$$

KL divergence:
How much more costly is it to encode data from P using Q?

$$\begin{aligned} D_{KL}(P \parallel Q) &= \sum_x P(x) log \frac{P(x)}{Q(x)} \\ &= H(P, Q) - H(P) \end{aligned} \tag{3}$$