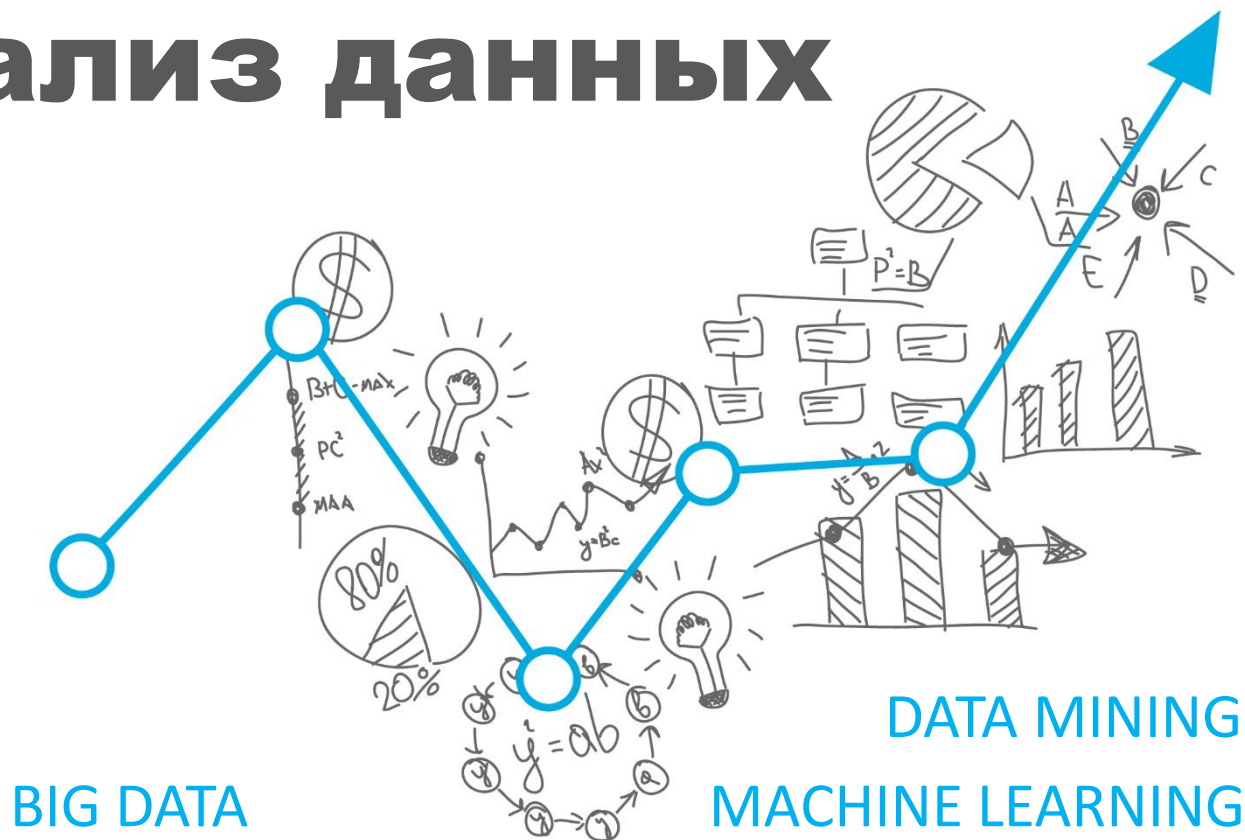
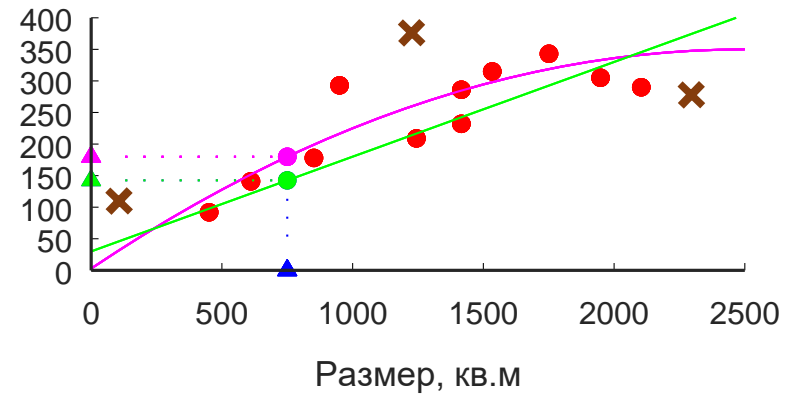


# Интеллектуальный анализ данных



## Лекция 7. Диагностика машинного обучения

№	Площадь	Комнат	Цена
1	2104	4	460
2	1416	5	232
3	1534	3	315
4	852	2	178
5	1948	5	305
6	950	4	293
7	611	2	141
8	1751	4	343
9	451	1	102
10	1244	2	209
11	1416	3	286
...	...	...	...



Построена модель регуляризованной линейной регрессии:

$$h_{\theta}(x) = \theta^T x \quad J(\theta) = \frac{1}{2m} \left[ \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^n \theta_j^2 \right]$$

$$J(\theta) \rightarrow \min \quad \Rightarrow \quad \theta^* - \text{оптимальный } \theta$$

Появляются новые данные, которые модель предсказывает плохо

Модель «плохо обобщается», «имеет плохую обобщающую способность»

**Что делать дальше?**

## ☐ Увеличить обучающий набор данных

Найти новые данные (измерения, опросы, эксперименты и т.д.)  
Может быть очень долго!

## ☐ Уменьшить число используемых признаков

( $x_1$  – площадь, ...  ~~$x_{11}$  – детская площадка, ...~~  
 $x_{100}$  – материал, ... )

## ☐ Увеличить число используемых признаков

Добавить:  $x_{101}$  – магазин рядом,  $x_{102}$  – лифт  
Опросы, эксперименты, ...

## ☐ Добавить полиномиальные признаки

$x_1, x_1^2, x_1x_2, x_1^2x_3^3, \dots$

## ☐ Увеличить параметр регуляризации $\lambda$

## ☐ Уменьшить параметр регуляризации $\lambda$

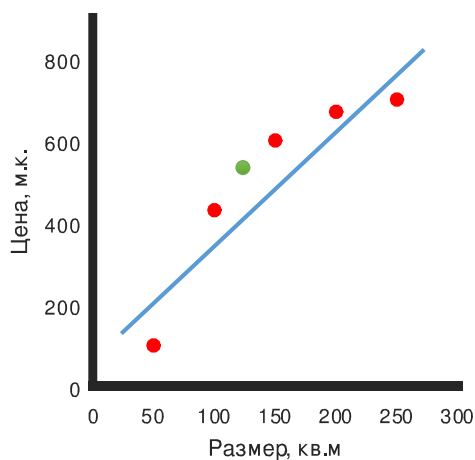
} Проверить легче всего,  
но тоже требует время  
и обоснованный выбор

Могут потребоваться часы, дни и месяцы!

Набор методик, которые позволяют:

- Оценить обобщающую способность модели
- Выявить недостатки модели
- Выделить перспективные и неперспективные подходы к повышению/улучшению способности

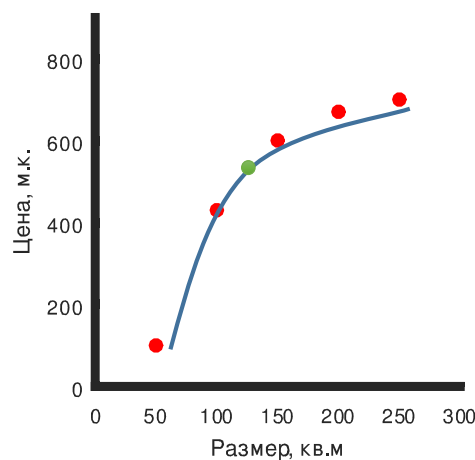
Диагностика тоже требует времени,  
но гораздо эффективнее,  
чем перебирать методики вслепую



$$h_{\theta}^{(1)}(x) = \theta_0 + \theta_1 x$$

Недообученная  
модель

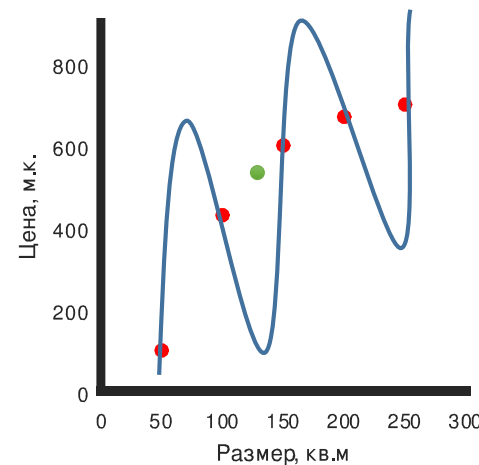
$$J^{(1)}(\theta) \rightarrow \min$$



$$h_{\theta}^{(2)}(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$

Подходящая модель

$$J^{(2)}(\theta) \rightarrow \min$$



$$h_{\theta}^{(3)}(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \dots + \theta_4 x^4$$

Переобученная  
модель

$$J^{(3)}(\theta) \rightarrow \min$$

$$J^{(1)}(\theta) > J^{(2)}(\theta) > J^{(3)}(\theta)$$

Все три модели минимальны на обучающем наборе данных

Как найти подходящую?

Обучающая выборка ~70%	№	Площадь	Комнат	Цена	
	1	2104	4	460	
	2	1416	5	232	
	3	1534	3	315	
	4	852	2	178	
	5	1948	5	305	
	6	950	4	293	
	7	611	2	141	
	8	1751	4	343	
	9	451	1	102	
	10	1244	2	209	
	11	1416	3	286	
	...	...	...	...	

$(x^{(1)}, y^{(1)})$   
 $(x^{(2)}, y^{(2)})$   
 $\vdots$   
 $(x^{(m)}, y^{(m)})$

Тестовая  
выборка  
~30%

$(x_{test}^{(1)}, y_{test}^{(1)})$   
 $(x_{test}^{(2)}, y_{test}^{(2)})$   
 $\vdots$   
 $(x_{test}^{(m_{test})}, y_{test}^{(m_{test})})$

Выбор должен  
быть случайным!

	№	Площадь	Комнат	Цена
Обучающая выборка	1	2104	4	460
	2	1416	5	232
	3	1534	3	315
	4	852	2	178
	5	1948	5	305
	6	950	4	293
	7	611	2	141
	8	1751	4	343
	9	451	1	102
	10	1244	2	209
	11	1416	3	286
	...	...	...	...
Тестовая выборка				

1) Обучаем модель на обучающей выборке, получаем  $\theta$

$$J(\theta) \rightarrow \min \Rightarrow \theta$$

2) Вычисляем функцию стоимости на тестовой выборке

$$J_{test}(\theta) = \frac{1}{2m_{test}} \sum_{i=1}^{m_{test}} \left( h_{\theta}(x_{test}^{(i)}) - y_{test}^{(i)} \right)^2$$

Обучающая выборка  $(x^{(1)}, y^{(1)})$ $(x^{(2)}, y^{(2)})$ $\vdots$ $(x^{(m)}, y^{(m)})$	№	Площадь	Комнат	Цена	Тестовая выборка  $(x_{test}^{(1)}, y_{test}^{(1)})$ $(x_{test}^{(2)}, y_{test}^{(2)})$ $\vdots$ $(x_{test}^{(m_{test})}, y_{test}^{(m_{test})})$
	1	2104	4	460	
	2	1416	5	232	
	3	1534	3	315	
	4	852	2	178	
	5	1948	5	305	
	6	950	4	293	
	7	611	2	141	
	8	1751	4	343	
	9	451	1	102	
	10	1244	2	209	
	11	1416	3	286	
	...	...	...	...	

Предположим, что модель линейной регрессии без регуляризации очень сильно переобучена. Какими будут значения функции стоимости на обучающей выборке ( $J(\theta)$ ) и на тестовой выборке ( $J_{test}(\theta)$ )?

- ☒  $J(\theta)$  будет мало, а  $J_{test}(\theta)$  будет велико
- ☐  $J(\theta)$  будет мало и  $J_{test}(\theta)$  будет мало
- ☐  $J(\theta)$  будет велико, а  $J_{test}(\theta)$  будет мало
- ☐  $J(\theta)$  будет велико и  $J_{test}(\theta)$  будет велико



**1) Вычисление ошибки тестовых данных для регрессии:**

$$J_{test}(\theta) = \frac{1}{2m_{test}} \sum_{i=1}^{m_{test}} \left( h_{\theta}(x_{test}^{(i)}) - y_{test}^{(i)} \right)^2$$

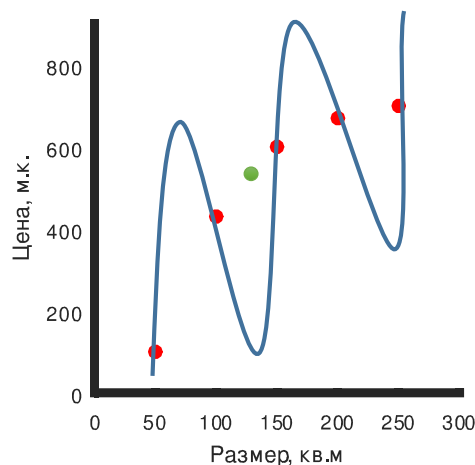
**2) Вычисление ошибки тестовых данных для классификации:**

$$J_{test}(\theta) = \frac{1}{2m_{test}} \sum_{i=1}^{m_{test}} err \left( h_{\theta}(x_{test}^{(i)}) - y_{test}^{(i)} \right)$$

$$err(h_{\theta}(x) - y) = \begin{cases} 1, \text{ если } (h_{\theta}(x) \geq 0.5 \text{ и } y = 0) \text{ или } (h_{\theta}(x) < 0.5 \text{ и } y = 1) \\ 0, \text{ в противном случае} \end{cases}$$

Ошибка классификации,  
т.е. число неверно классифицированных данных

## Зачем нужен отдельный тестовый набор данных?



Если подобрать параметры  $\theta$  на обучающем наборе данных, то оценивать качество этого обучения на том же наборе данных — плохая идея!

Оценка  $J(\theta)$  на обучающем наборе будет меньше, чем обобщенная оценка.

Оценить, насколько хорошо обобщается модель (как она работает на данных, которых не было в обучении) можно только на отдельном наборе данных.

Если мы подбираем параметры модели, коэффициент регуляризации, степень полинома и т. д. на каком то наборе данных, то оценить результаты мы сможем лишь на отдельном наборе.

Какую степень полинома выбрать?

$d = 1$	$h_{\theta}(x) = \theta_0 + \theta_1 x$	$\rightarrow \theta^{(1)}$	$\rightarrow J_{test}(\theta^{(1)})$	$\min_d J_{test}$
$d = 2$	$h_{\theta}(x) = \theta_0 + \theta_1 x + \theta_2 x^2$	$\rightarrow \theta^{(2)}$	$\rightarrow J_{test}(\theta^{(2)})$	
$d = 3$	$h_{\theta}(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$	$\rightarrow \theta^{(3)}$	$\rightarrow J_{test}(\theta^{(3)})$	
$\vdots$				
$d = 5$	$h_{\theta}(x) = \theta_0 + \theta_1 x + \dots + \theta_5 x^5$	$\rightarrow \theta^{(5)}$	$\rightarrow J_{test}(\theta^{(5)})$	
$\vdots$				
$d = 10$	$h_{\theta}(x) = \theta_0 + \theta_1 x + \dots + \theta_{10} x^{10}$	$\rightarrow \theta^{(10)}$	$\rightarrow J_{test}(\theta^{(10)})$	

Добавим параметр  $d$  – степень полинома модели.

$$h_{\theta}(x) = \theta_0 + \theta_1 x + \dots + \theta_5 x^5$$

Насколько хорошо обобщается эта модель? Можем ли мы по  $J_{test}(\theta^{(5)})$  оценивать обобщающую способность?

(в сравнении с другими решениями: нейросеть, другой набор функций)

Не можем! Параметр  $d$  был подобран по тестовой выборке данных, а значит оценка качества его выбора не будет обобщенной.

Обучающая выборка ~60%	№	Площадь	Комнат	Цена	
	1	2104	4	460	
	2	1416	5	232	
	3	1534	3	315	
	4	852	2	178	
	5	1948	5	305	
	6	950	4	293	
	7	611	2	141	
	8	1751	4	343	
	9	451	1	102	
	10	1244	2	209	
	11	1416	3	286	
	...	...	...	...	

$$(x_{cv}^{(1)}, y_{cv}^{(1)})$$

$$(x_{cv}^{(2)}, y_{cv}^{(2)})$$

⋮

$$(x_{cv}^{(m_{cv})}, y_{cv}^{(m_{cv})})$$

Валидационная  
выборка, ~20%

Тестовая выборка  
~20%

Валидационная (проверочная) выборка:

CV (cross validation) – перекрестная проверка

$(x, y)$  – обучение модели (коэффициенты полинома,  
веса нейронной сети, ...)

$(x_{cv}, y_{cv})$  – подбор параметров (степень полинома,  
регуляризация, количество слоев/нейронов, ...)

$(x_{test}, y_{test})$  – оценка обобщающей способности модели

$$(x_{test}^{(1)}, y_{test}^{(1)})$$

$$(x_{test}^{(2)}, y_{test}^{(2)})$$

⋮

$$(x_{test}^{(m_{test})}, y_{test}^{(m_{test})})$$

**Выбор должен  
быть случайным!**

1) Ошибка на обучении:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m \left( h_{\theta}(x^{(i)}) - y^{(i)} \right)^2$$

2) Ошибка на валидации:

$$J_{cv}(\theta) = \frac{1}{2m_{cv}} \sum_{i=1}^{m_{cv}} \left( h_{\theta}(x_{cv}^{(i)}) - y_{cv}^{(i)} \right)^2$$

3) Ошибка на тесте:

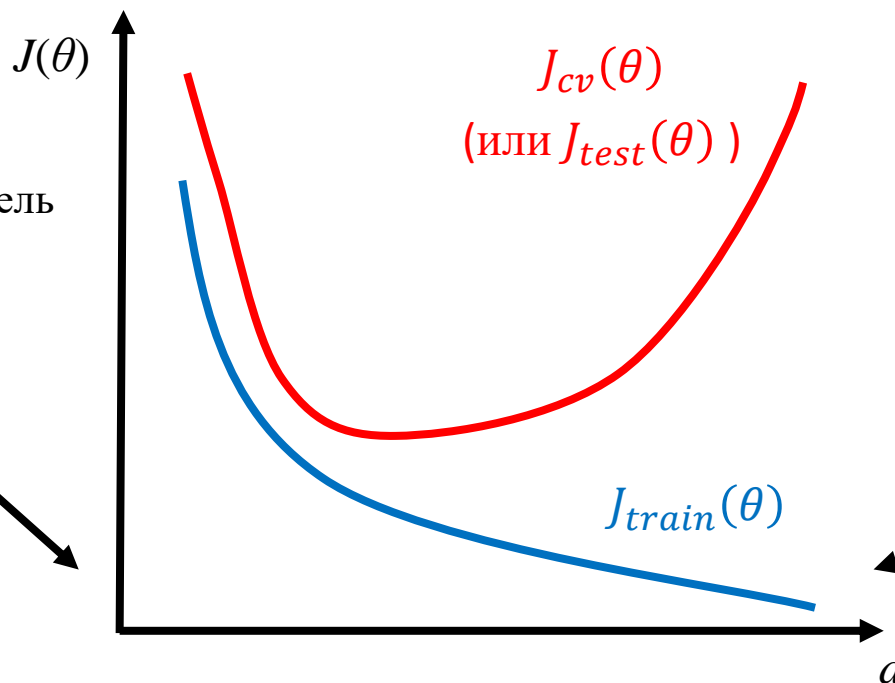
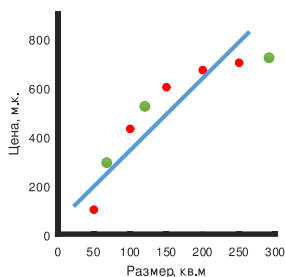
$$J_{test}(\theta) = \frac{1}{2m_{test}} \sum_{i=1}^{m_{test}} \left( h_{\theta}(x_{test}^{(i)}) - y_{test}^{(i)} \right)^2$$

Вы построили полиномиальную регрессионную модель, выбрали степень полинома  $d$ , и оценили обобщающую способность полученной модели. При этом обнаружили, что значение ошибки  $J_{cv}(\theta)$  меньше, чем  $J_{test}(\theta)$ .

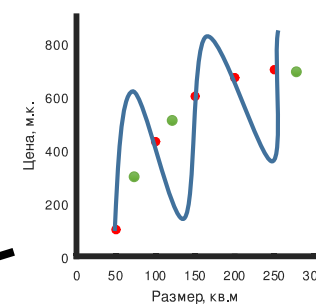
По какой причине это происходит?

- ☒ Потому что параметр  $d$  подобран по тестовой выборке
- ☒ Потому что параметр  $d$  подобран по валидационной выборке
- ☒ Потому что валидационная выборка обычно меньше тестовой
- ☒ Потому что валидационная выборка обычно больше тестовой

Недообученная модель



Переобученная модель



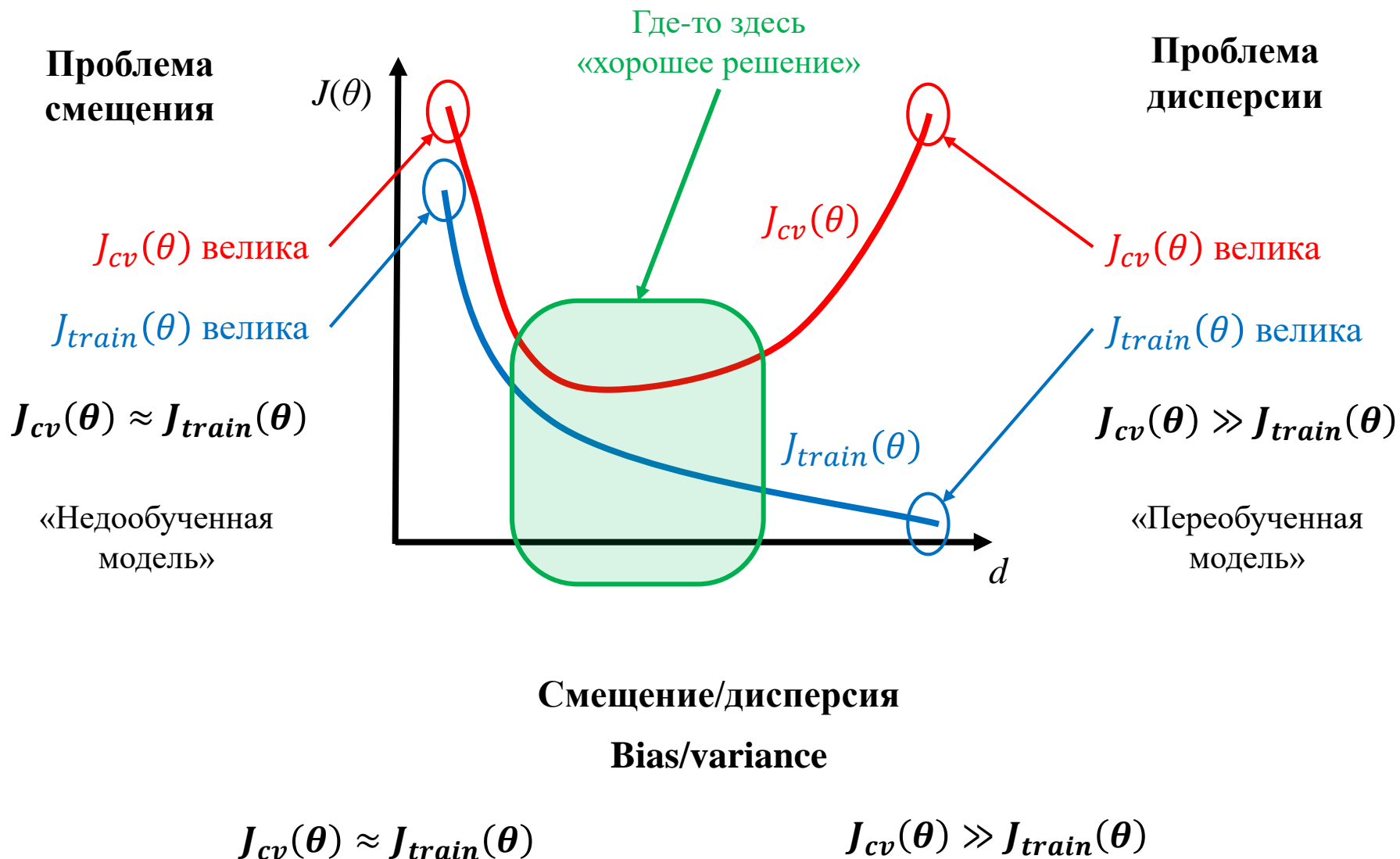
Ошибка на обучении:

$$J_{train}(\theta) = \frac{1}{2m} \sum_{i=1}^m \left( h_{\theta}(x^{(i)}) - y^{(i)} \right)^2$$

Ошибка на валидации или на тесте:

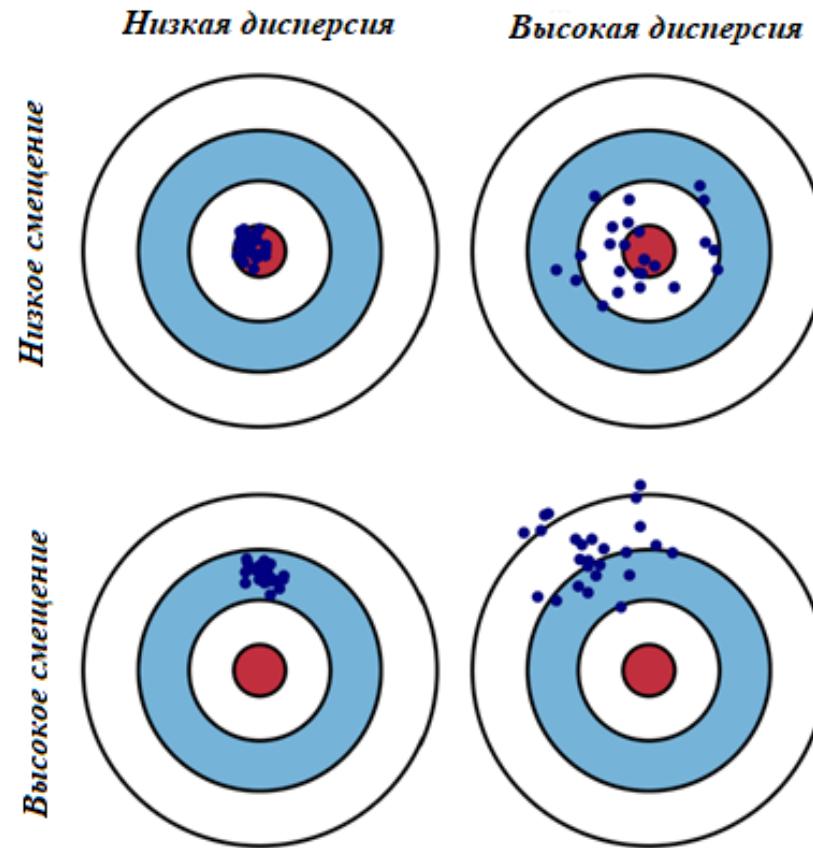
$$J_{cv}(\theta) = \frac{1}{2m_{cv}} \sum_{i=1}^{m_{cv}} \left( h_{\theta}(x_{cv}^{(i)}) - y_{cv}^{(i)} \right)^2$$

$$J_{test}(\theta) = \frac{1}{2m_{test}} \sum_{i=1}^{m_{test}} \left( h_{\theta}(x_{test}^{(i)}) - y_{test}^{(i)} \right)^2$$





Цель машинного обучения – получить хороший алгоритм



Bias –  
базис, основа,  
точка отчёта

Variance –  
вариабельность

Смещение – неверный алгоритм, не способный найти решение

Дисперсия – слишком вариативный алгоритм, скорее «промахнётся»

Вы решаете задачу классификации. В процессе обучения вы получили ошибку классификации на **обучающем** наборе, равную **0.1**, и на **валидационном** наборе, равную **0.3**. С какой проблемой вероятнее всего вы столкнулись?

☒ Смещения (переобучения)

☒ Смещения (недообучения)

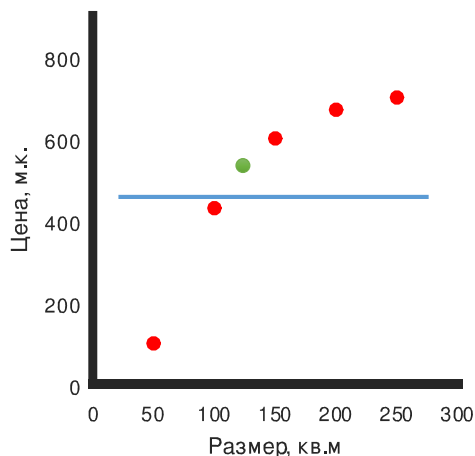
☒ Дисперсии (переобучения)

☒ Дисперсии (недообучения)

Модель:  $h_{\theta}(x) = \theta_0 + \cancel{\theta_1 x} + \cancel{\theta_2 x^2} + \cancel{\theta_3 x^3} + \cancel{\theta_4 x^4}$

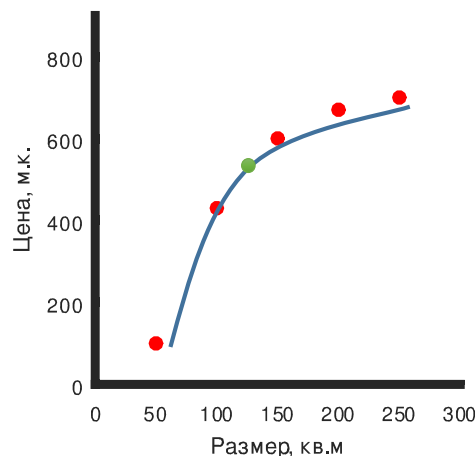
$$J(\theta) = \frac{1}{2m} \left[ \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^n \theta_j^2 \right]$$

Регуляризация



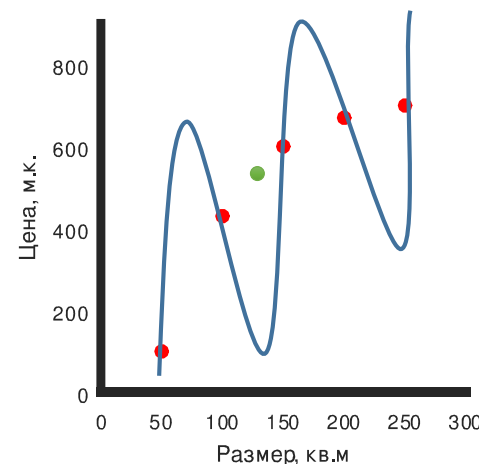
Большое  $\lambda$

Большое смещение  
(недообучение)



Какое  $\lambda$  ?


Подходящая  
модель



Малое  $\lambda$

Высокая дисперсия  
(переобучение)

Модель:  $h_{\theta}^{(3)}(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

$$J(\theta) = \frac{1}{2m} \left[ \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^n \theta_j^2 \right]$$


1) Ошибка на обучении:

$$J_{train}(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

2) Ошибка на валидации:

$$J_{cv}(\theta) = \frac{1}{2m_{cv}} \sum_{i=1}^{m_{cv}} (h_{\theta}(x_{cv}^{(i)}) - y_{cv}^{(i)})^2$$

3) Ошибка на тесте:

$$J_{test}(\theta) = \frac{1}{2m_{test}} \sum_{i=1}^{m_{test}} (h_{\theta}(x_{test}^{(i)}) - y_{test}^{(i)})^2$$

С регуляризацией

Без  
регуляризации!

Выбор значения параметра регуляризации  $\lambda$

1. Возьмем $\lambda = 0$	$\theta^{(1)} = \min_{\theta} J(\theta   \lambda = 0)$	$J_{cv}(\theta^{(1)})$
2. Возьмем $\lambda = 0.01$	$\theta^{(2)} = \min_{\theta} J(\theta   \lambda = 0.01)$	$J_{cv}(\theta^{(2)})$
3. Возьмем $\lambda = 0.02$	$\theta^{(3)} = \min_{\theta} J(\theta   \lambda = 0.02)$	$J_{cv}(\theta^{(3)})$
4. Возьмем $\lambda = 0.04$	$\theta^{(4)} = \min_{\theta} J(\theta   \lambda = 0.04)$	$J_{cv}(\theta^{(4)})$
5. Возьмем $\lambda = 0.08$	$\theta^{(5)} = \min_{\theta} J(\theta   \lambda = 0.08)$	$J_{cv}(\theta^{(5)})$
$\vdots$		
12. Возьмем $\lambda = 10$	$\theta^{(12)} = \min_{\theta} J(\theta   \lambda = 10)$	$J_{cv}(\theta^{(12)})$

Наилучшим значением для  $\lambda$  будет то,  
которое соответствует  $\min_i J_{cv}(\theta^{(i)})$

$J_{test}(\theta^{(4)})$  – оценка обобщающей способности модели

Функция стоимости:

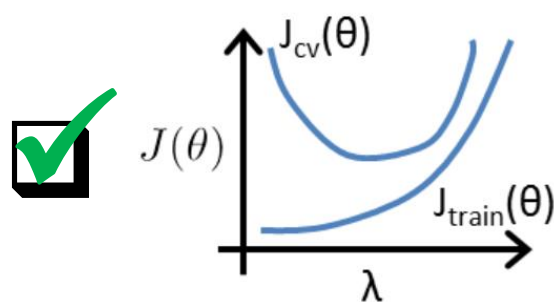
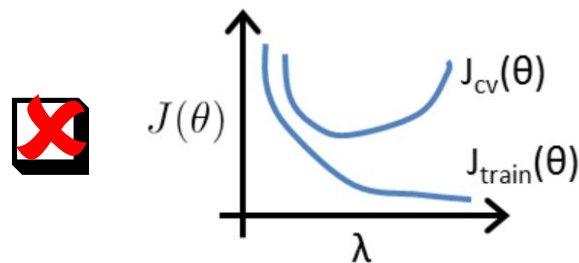
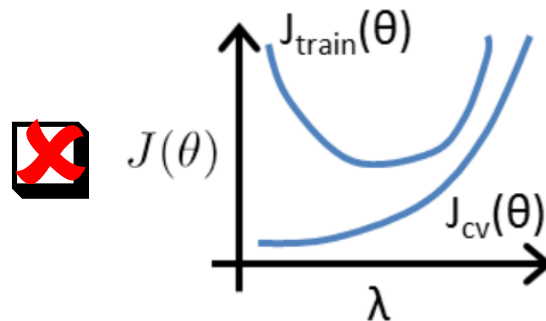
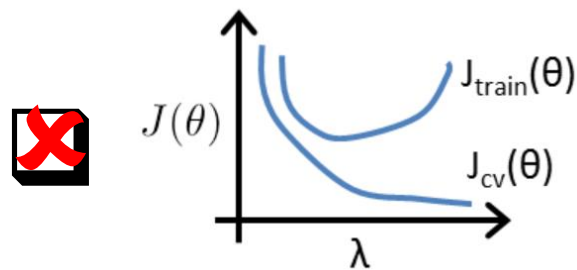
$$J(\theta) = \frac{1}{2m} \left[ \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^n \theta_j^2 \right]$$

Ошибка на обучении:

$$J_{train}(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

Ошибка на валидации:

$$J_{cv}(\theta) = \frac{1}{2m_{cv}} \sum_{i=1}^{m_{cv}} (h_{\theta}(x_{cv}^{(i)}) - y_{cv}^{(i)})^2$$



Функция стоимости:

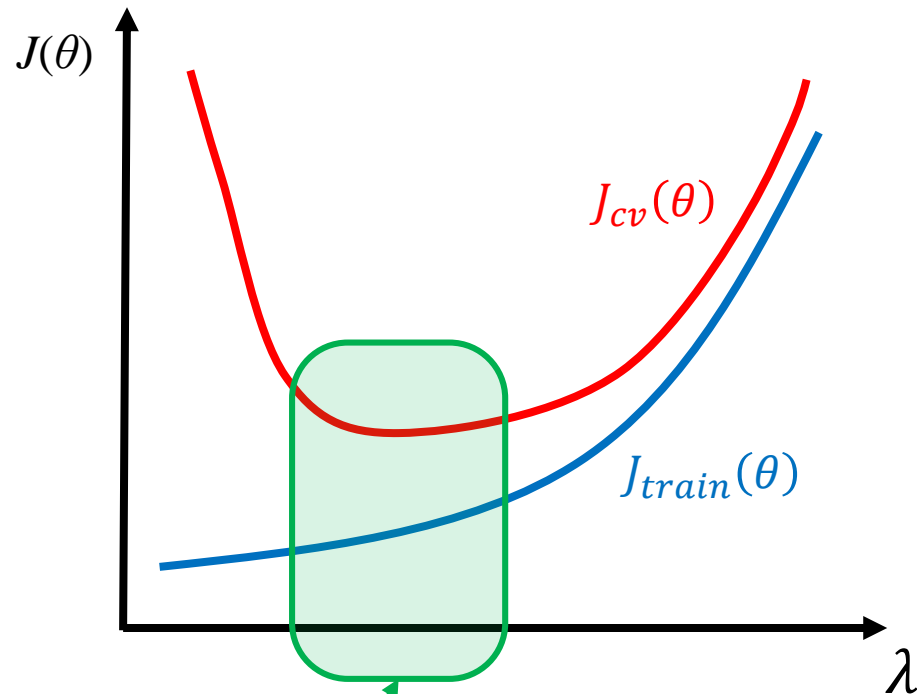
$$J(\theta) = \frac{1}{2m} \left[ \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^n \theta_j^2 \right]$$

Ошибка на обучении:

$$J_{train}(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

Ошибка на валидации:

$$J_{cv}(\theta) = \frac{1}{2m_{cv}} \sum_{i=1}^{m_{cv}} (h_{\theta}(x_{cv}^{(i)}) - y_{cv}^{(i)})^2$$



Где-то здесь  
«хорошее решение»

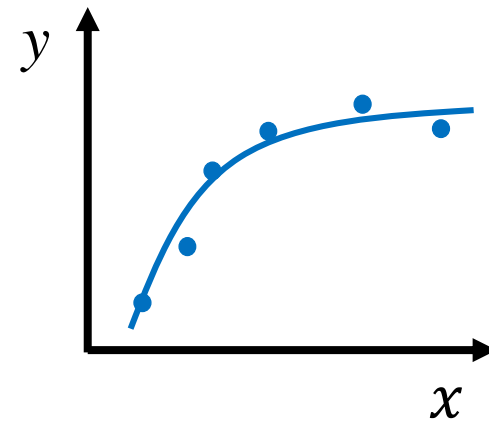
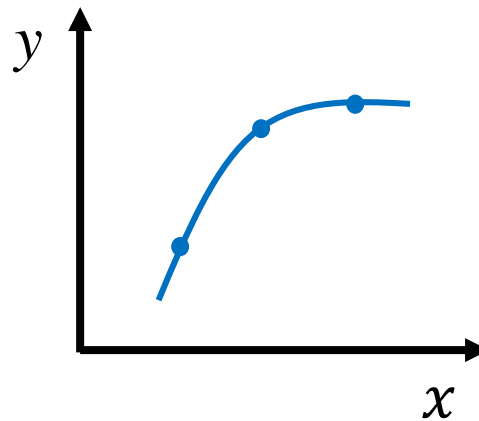
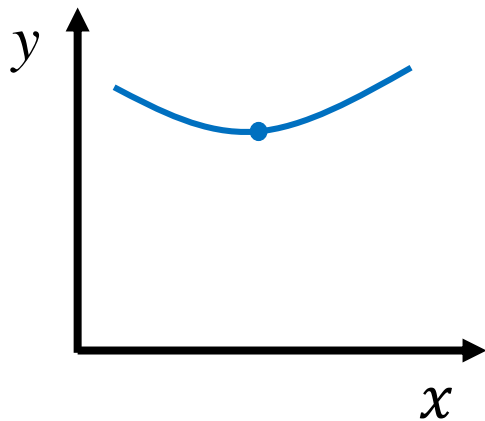
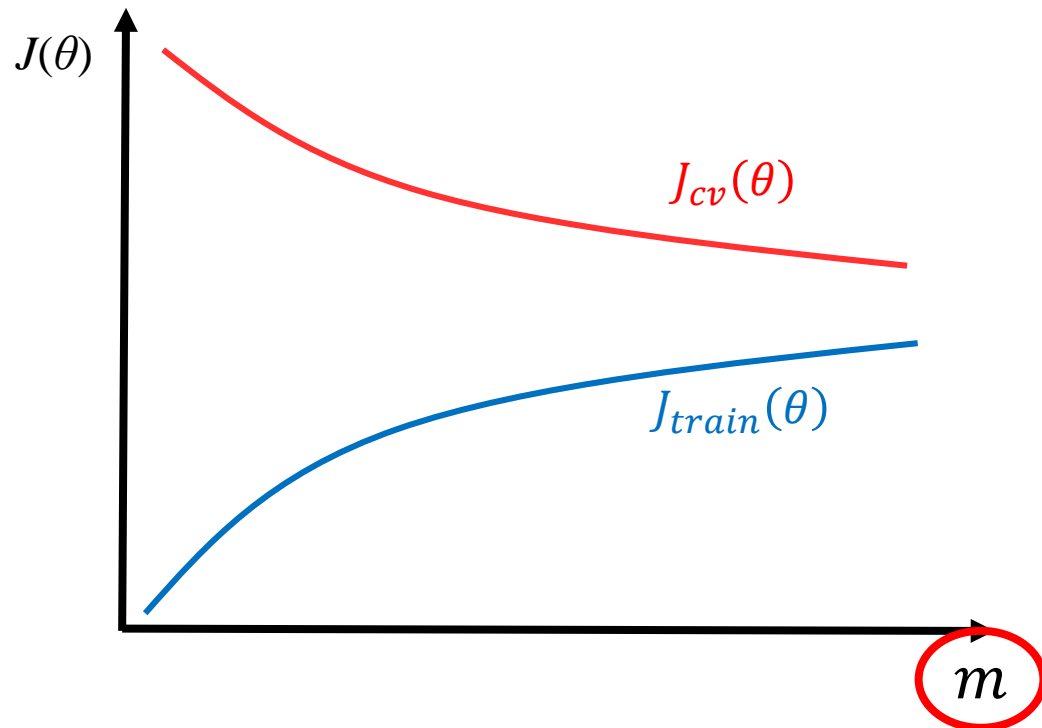
Ошибка на обучении:

$$J_{train}(\theta) = \frac{1}{2m} \sum_{i=1}^m \left( h_{\theta}(x^{(i)}) - y^{(i)} \right)^2$$

Ошибка на валидации:

$$J_{cv}(\theta) = \frac{1}{2m_{cv}} \sum_{i=1}^{m_{cv}} \left( h_{\theta}(x_{cv}^{(i)}) - y_{cv}^{(i)} \right)^2$$

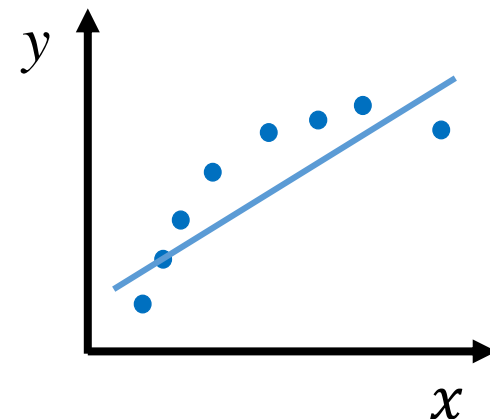
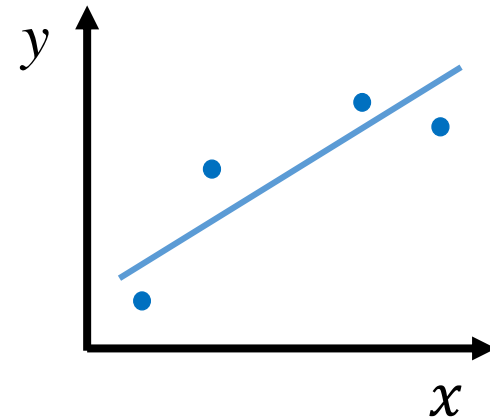
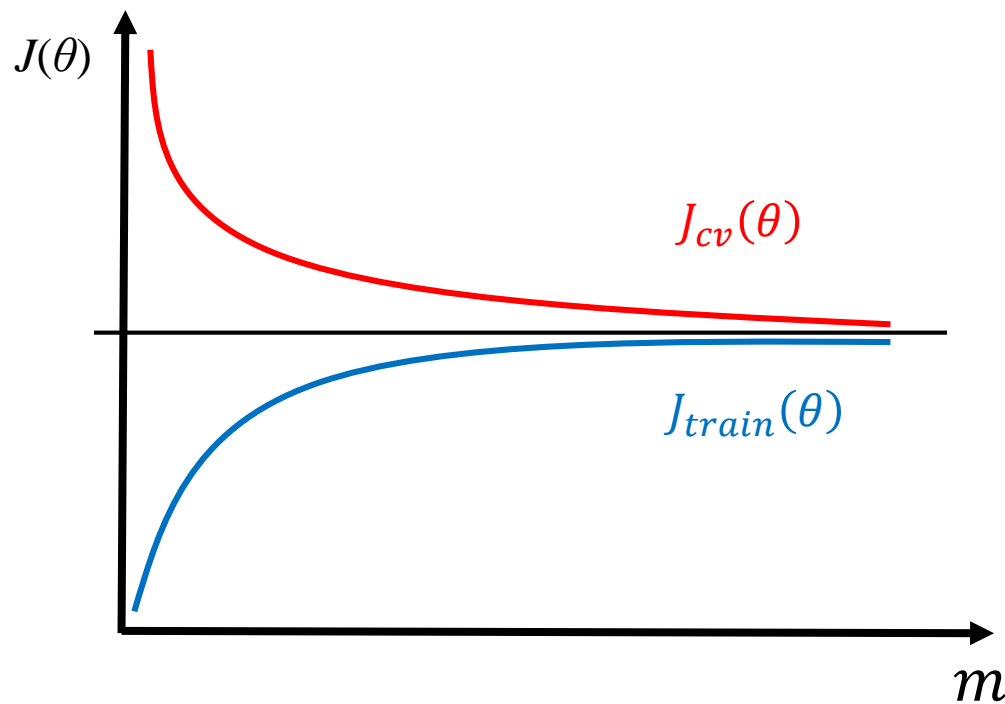
$$h_{\theta}(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$





Диагностика проблемы смещения

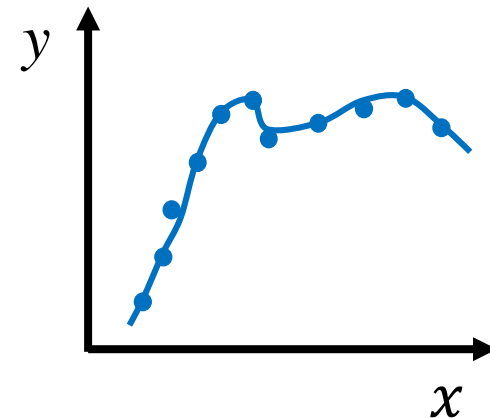
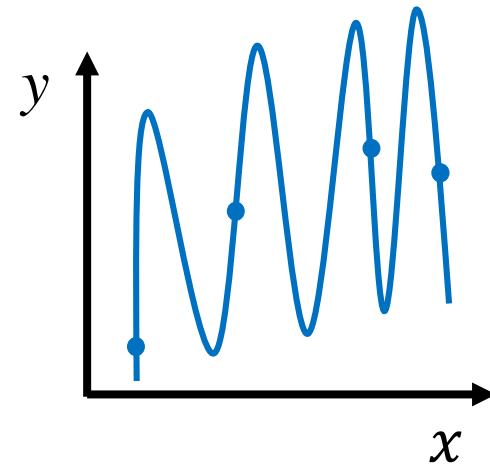
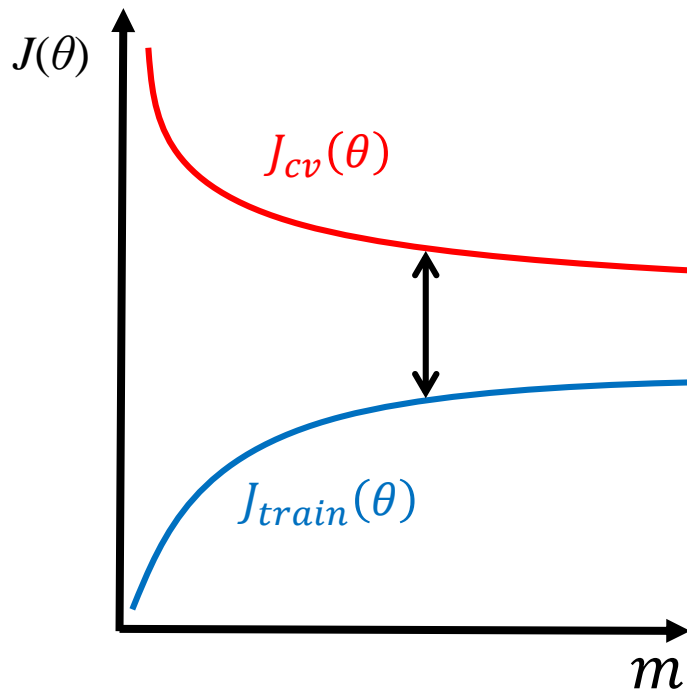
$$h_{\theta}(x) = \theta_0 + \theta_1 x$$



Добавление новых данных не улучшает решение!  
Но увеличение сложности модели улучшит!

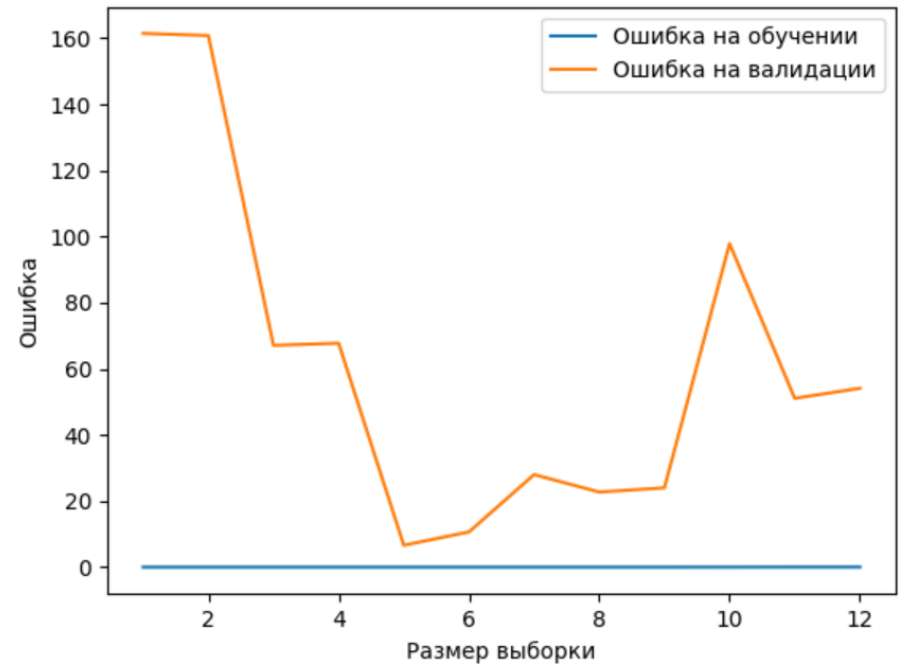
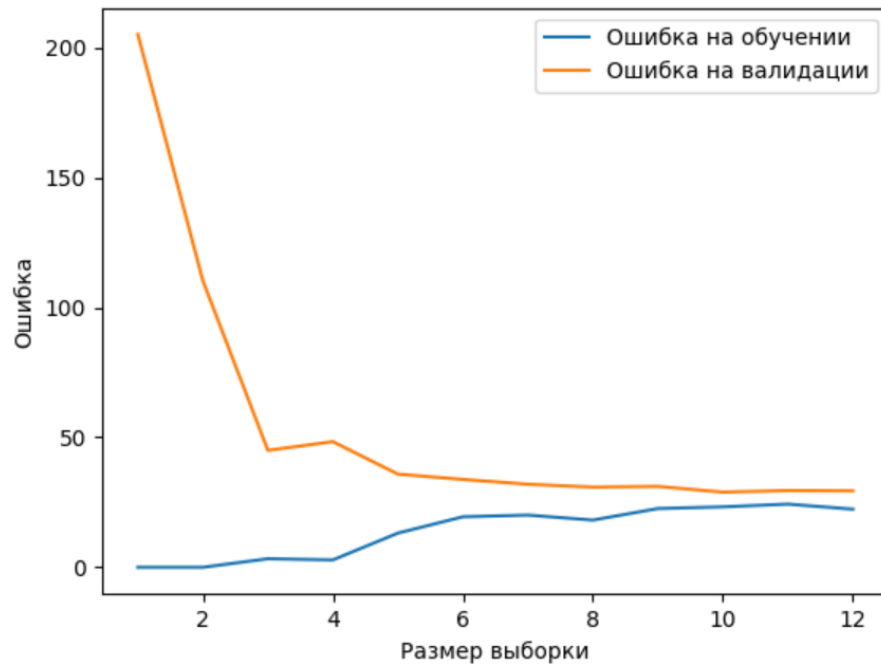
Диагностика проблемы дисперсии

$$h_{\theta}(x) = \theta_0 + \theta_1 x + \dots + \theta_{100} x^{100}$$



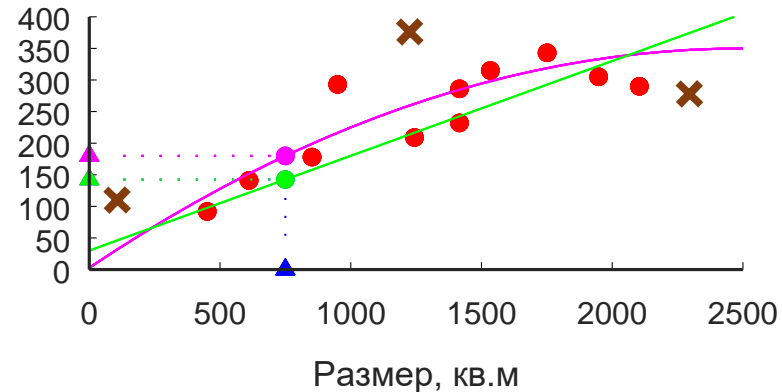
Добавление новых данных помогает улучшить решение!  
Увеличение сложности модели не поможет!

Реальные кривые обучения могут быть зашумленными и неровными



№	Площадь	Комнат	Цена
1	2104	4	460
2	1416	5	232
3	1534	3	315
4	852	2	178
5	1948	5	305
6	950	4	293
7	611	2	141
8	1751	4	343
9	451	1	102
10	1244	2	209
11	1416	3	286
...	...	...	...

Вернемся к исходному вопросу:



Построена модель регуляризованной линейной регрессии:

$$h_{\theta}(x) = \theta^T x \quad J(\theta) = \frac{1}{2m} \left[ \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^n \theta_j^2 \right]$$

$$J(\theta) \rightarrow \min \quad \Rightarrow \quad \theta^* - \text{оптимальный } \theta$$

Появляются новые данные, которые модель предсказывает плохо

Модель «плохо обобщается», «имеет плохую обобщающую способность»

Что делать дальше?

Выполнить диагностику решения,  
построить кривые обучения, найти проблемы

Далее варианты:

- ☐ Увеличить обучающий набор данных

Помогает, если имеется проблема дисперсии

- ☐ Уменьшить число используемых признаков

Помогает, если имеется проблема дисперсии

- ☐ Увеличить число используемых признаков

Помогает, если имеется проблема смещения

- ☐ Добавить полиномиальные признаки

Помогает, если имеется проблема смещения

- ☐ Увеличить параметр регуляризации  $\lambda$

Помогает, если имеется проблема дисперсии

- ☐ Уменьшить параметр регуляризации  $\lambda$

Помогает, если имеется проблема смещения

## Небольшая нейронная сеть

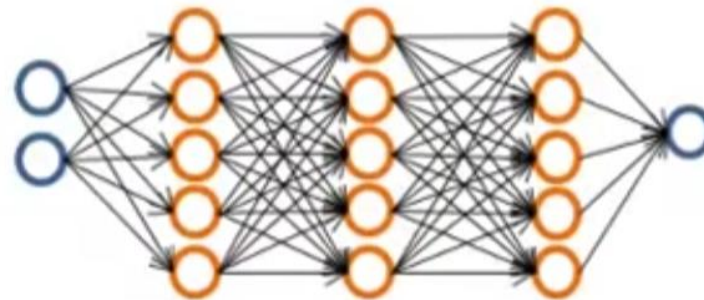


- +** Требуется меньше памяти
- +** Быстрее обучается и работает
- Решает простые задачи
- Возможна проблема смещения

Решение:

- Усложнение сети, добавление слоев и нейронов

## Большая нейронная сеть



- +** Решение сложных задач
- Требуется больше памяти
- Долго вычисляется
- Возможна проблема дисперсии

Решение:

- Упрощение сети
- Добавление регуляризации (предпочтительное решение)

Вы обучаете нейронную сеть с одним скрытым слоем. В процессе обучения вы определили, что ошибка на **валидационном** наборе сильно больше, чем на **обучающем** наборе.

Может ли вам помочь в этой ситуации увеличение числа скрытых слоев нейронной сети?

- ☐ Да, поскольку это увеличивает число параметров и позволяет представить более сложную функцию
- ☐ Да, поскольку имеется проблема смещения
- ☐ Нет, поскольку имеется проблема смещения
- ☒ Нет, поскольку имеется проблема дисперсии