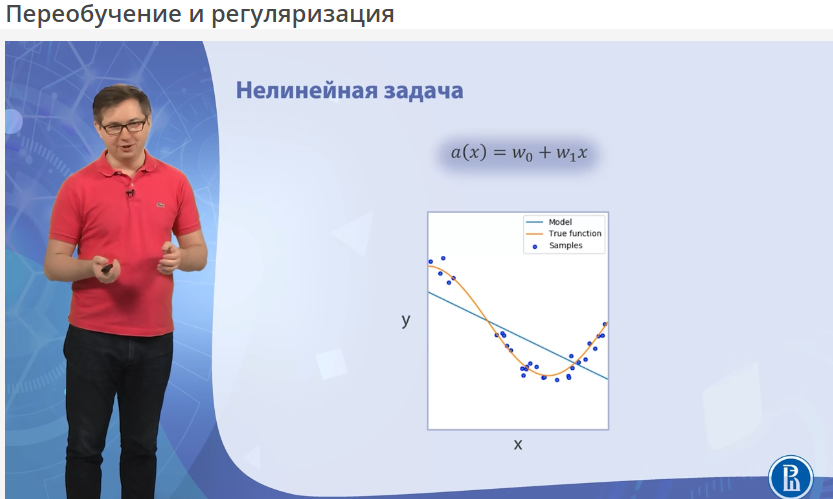
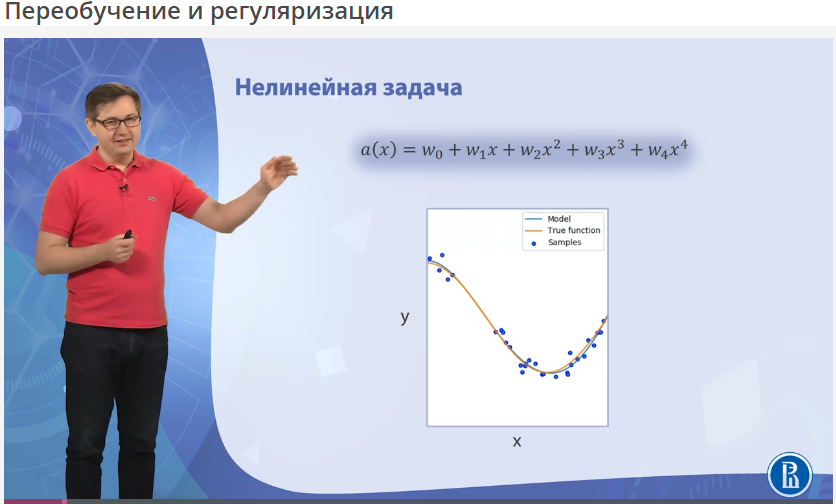


Что такое переобучение линейных моделей и как с этим бороться



Линейная модель может плохо отражать ситуацию

Если сделать полиномиальную модель

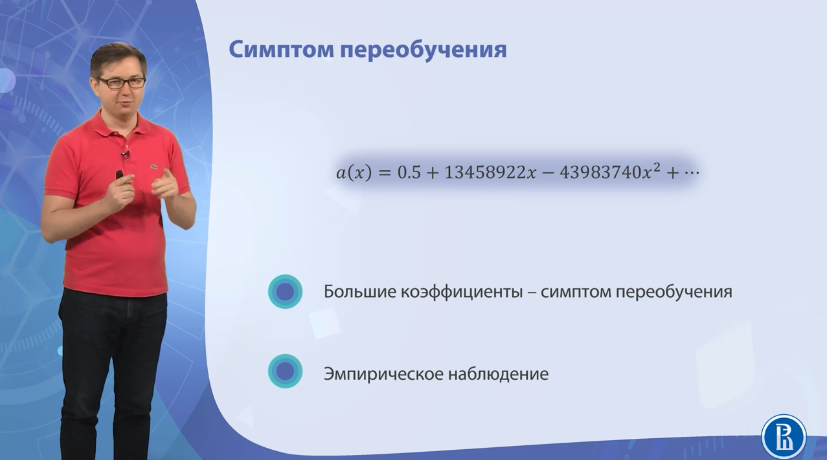


Модель почти идеальная, но проходит не через все точки. Увеличиваем число степеней….

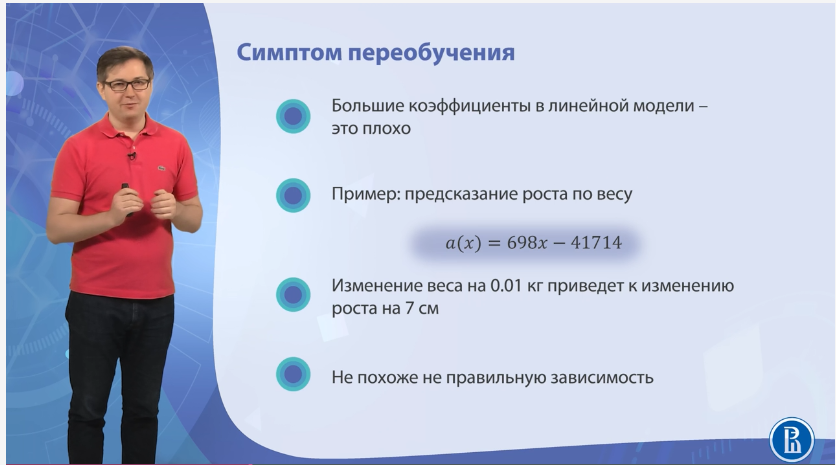


Упс

Ошибка на обучающей модели будет меньше, но на новых данных все будет ужасно – модель ПЕРЕОБУЧЕНА

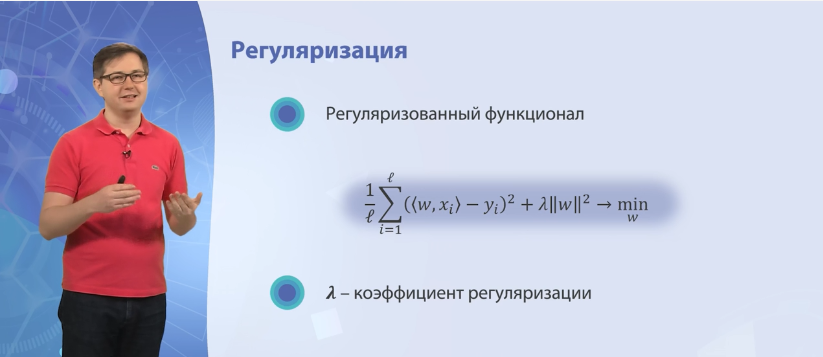


При переобучении веса при признаках огромные по модулю (сильно отличаются от нуля) – это симптом.

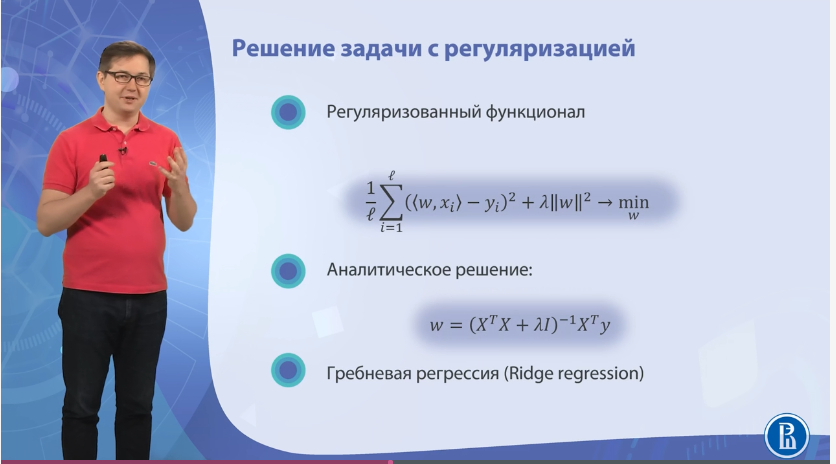




Регуляризатор – штраф модели – квадрат евклидовой нормы весов

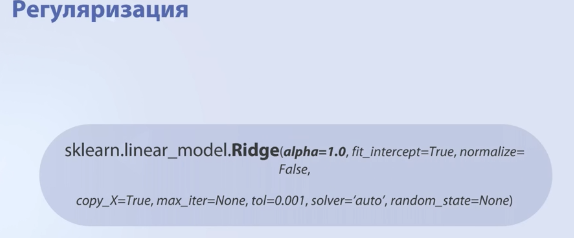


Лямбда говорит о том, насколько нам важно чтобы веса были небольшие. Если лямбда огромная – регуляризация будет сильной.

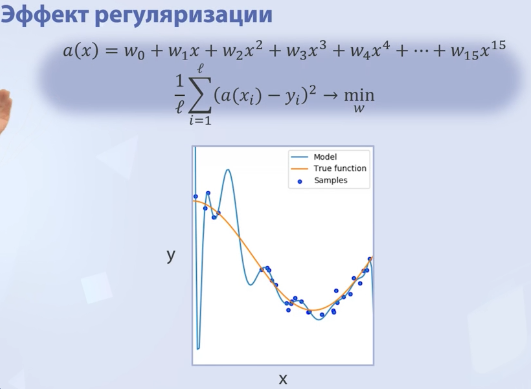


I – диагональная матрица, лямбды стоят на диагонали

Такое добавление диагональной матрицы к ковариационной может даже вырожденную матрицу сделать невырожденной



Alpha- коэффициент регуляризации









при большом коэффициенте регуляризации модели выгоднее иметь большие веса, чем пытаться уменьшить MSE

На обучающей выборке коэффициенте регуляризации выгоднее иметь нулевым, а подбирать его уже на отложенной выборке или кросс-валидации

В коэффициент регуляризации нельзя включать свободный коэффициент w0 – то модель не сможет обучиться



Можно использовать не евклидову норму весов, а Лассо – сумму весов. Использование его приводит к тому, что некоторые веса обнуляются, использовать те признаки, которые реально важны.



L1 – использует модули весов, а L2 – евклидовы нормы

