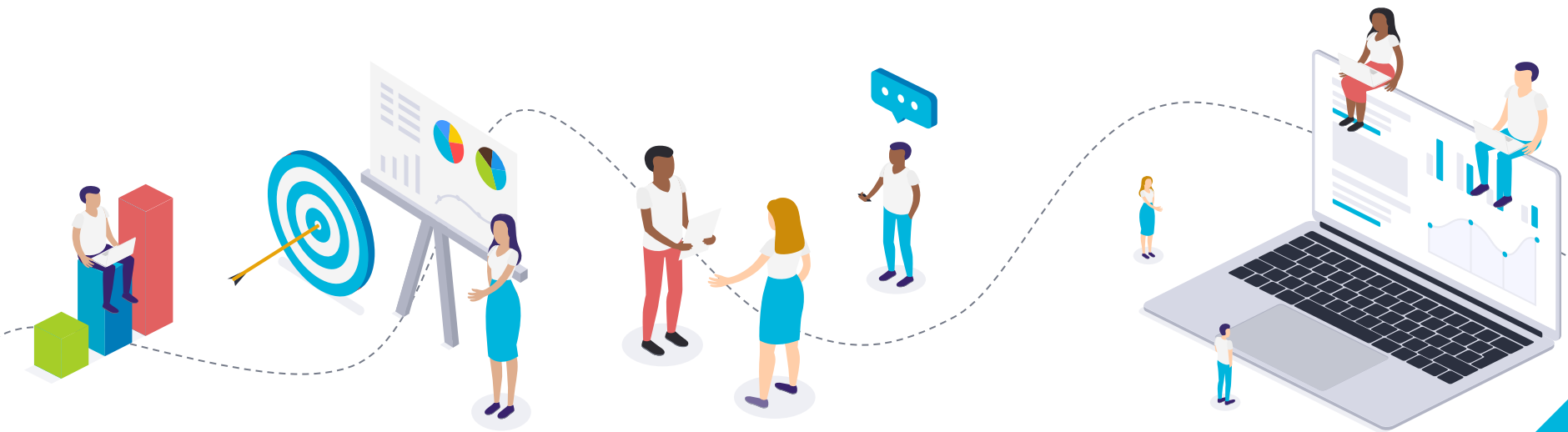
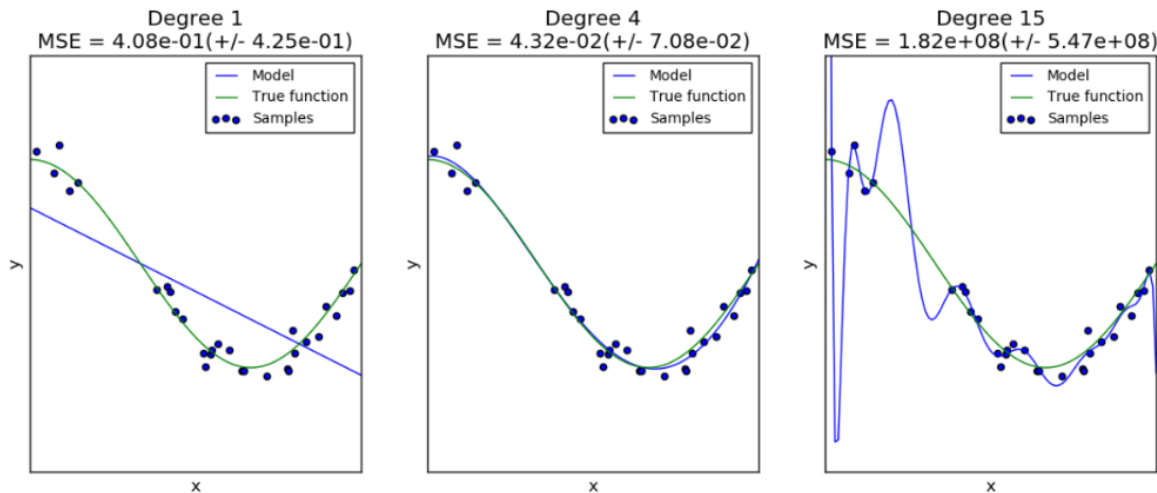


Программа профессиональной переподготовки «Технологии искусственного интеллекта, визуализации и анализа данных»



Переобучение и недообучение

- Модели с очень низкой сложностью являются **недообученными**, такие модели будут плохо предсказывать ответы как для обучающей выборки, так и для новых данных.
- Модели с очень высокой сложностью являются **переобученными**, то есть будут очень хорошо предсказывать ответы обучающей выборки, но на новых данных будут работать очень плохо.



Переобучение

Чтобы выявить переобучение модели – нужно оценить ее качество на новых данных!

- Отложенная выборка
 - Перекрестная проверка
-
- «Симптом» переобучения линейных моделей – это большие веса.
 - Другая ситуация, из-за которой возникает переобучение – *мультиколлинеарность*. Так называется проблема, при которой признаки в выборке являются линейно зависимыми.

Регуляризация – добавление дополнительных ограничений

$$Q(\omega, X) + \lambda R(\omega) \rightarrow \min_{\omega}$$

↑ ↑
Коэффициент регуляризации регуляризатор

Виды регуляризаторов:

- $L_2: R(\omega) = \sum_{j=1}^d \omega_j^2$ (квадратичный регуляризатор)
- $L_1: R(\omega) = \sum_{j=1}^d |\omega_j|$

L_1 обладает интересным свойством: если применять такой регуляризатор, некоторые веса оказываются равными нулю, то есть L_1 -регуляризатор производит отбор признаков и позволяет использовать в модели не все признаки, а только самые важные из них.

Регуляризованная линейная регрессия

- **Гребневая регрессия** – это модель линейной регрессии с L_2 регуляризацией.
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html
- **Метод Lasso** (оператор наименьшего абсолютного стягивания и отбора, least absolute shrinkage and selection operator) – это модель линейной регрессии с L_1 регуляризацией.
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html
- **Эластичная сеть** (ElasticNet) позволяет варьировать соотношение L_1 к L_2 .
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html

Параметр регуляризации λ – это гиперпараметр, который задается при создании экземпляра класса с помощью параметра **alpha**.

В установить ElasticNet если установить параметр **l1_ratio** равным 1.0, то регрессор ElasticNet будет равен Lasso.

Гиперпараметры. Как выбирать значения гиперпараметров?

Гиперпараметры – параметры алгоритмов, значения которых устанавливаются перед запуском процесса обучения.

- Степень полинома (*degree*)
- Параметр регуляризации λ

Как выбирать значения гиперпараметров?

Можно разделить выборку на части:

1 вариант:

Обучающая

Для обучения

Валидационная

Для оценки моделей на данных, которые не участвовали в обучении

Контрольная

Для проверки адекватности модели на новых данных

2 вариант:

Для перекрестной проверки

Контрольная