1. Тема вебинара
2. Представиться
3. Структура вебинара.

   Проговорить темы;
   обозначить время;
   сказать, чтоб вопросы задавали голосом, также предупредить, что буду спрашивать персонально случайно выбранного слушателя, для большего вовлечения.

4. Машинное обучение.

   Вспомним какие виды машинного обучения вообще существуют.
   Условно МЛ можно разделить на 4 вида: Классическое МЛ,Ансамблевые методы, Нейронные сети/глубокое обучение DL, Обучение с подкреплением. Со всеми этими видами МЛ вы постепенно ознакомитесь на курсе. Сейчас мы с вами начнем изучение с КМЛ с учителем, а именно классификации.
   Вспомним, что обучение с учителем используется всякий раз, когда мы хотим предсказать определенный результат (ответ) по какому-либо объекту, и у нас есть пары объект-ответ. Мы строим модель машинного обучения на основе этих пар объект-ответ, которые и составляют нашу обучающую выборку. Наша же цель состоит в том, чтобы получить точные прогнозы для новых, никогда ранее не встречавшихся данных т.е., на основе взаимосвязи, которую выявит алгоритм предсказывать значения/ ответы  для новых объектов.

5. Задачи классификации

**Цель классификации** состоит в том, чтобы спрогнозировать метку класса (class label), которая представляет собой выбор из заранее определенного списка возможных вариантов.

Классификация разделяется на:

- **бинарную классификацию** (binary classification), которая является частным случаем разделения на два класса;
- **мультиклассовую классификацию** (multiclass classification), когда в классификации участвует более двух классов.

Бинарную классификацию можно представить как попытку ответить на поставленный вопрос в формате «да/нет». Кредитный скоринг - дать кредит или нет. При распознавании образов - кошка или собака.

Но если мы хотим определить например породу собаки, то тут вариантов гораздо больше, чем два, следовательно. имеем дело с мультиклассовой

классификацией.

## 6. Линейные модели для классификации

Формула очень похожа на формулу линейной регрессии, но теперь вместо того, чтобы просто возвратить взвешенную сумму признаков, мы задаем для прогнозируемого значения порог, равный нулю.

Если функция меньше нуля, мы прогнозируем класс 0, если она больше нуля, мы прогнозируем класс 1.

**(Бинарный) линейный классификатор** – это классификатор, который разделяет два класса с помощью линии, плоскости или гиперплоскости.

Двумя наиболее распространенными алгоритмами линейной классификации являются логистическая регрессия (logistic regression), реализованная в классе linear_model.LogisticRegression, и линейный метод опорных векторов (linear support vector machines) или линейный SVM, реализованный в классе svm.LinearSVC (SVC расшифровывается как support vector classifier – классификатор опорных векторов). Несмотря на свое название, логистическая регрессия является алгоритмом классификации, а не алгоритмом регрессии, и его не следует путать с линейной регрессией.

## 7. Обучение ЛК

В задаче регрессии имеется непрерывность возможных ответов, и при таких условиях достаточно странно требовать полного совпадения ответов модели и истинных ответов  гораздо логичнее говорить об их близости. Способов посчитать близость двух чисел (прогноза и истинного ответа) достаточно много, и поэтому при обсуждении регрессии у нас возникло большое количество функционалов ошибки.
В случае с бинарной классификацией всё гораздо проще: у нас всего два возможных ответа алгоритма и, очевидно, мы хотим видеть как можно больше правильных ответов.

Этот функционал является дискретным (конечным) относительно весов, и поэтому искать его минимум с помощью градиентных методов мы не сможем.
Доля правильных ответов - точность:
Данная метрика, однако, имеет существенный недостаток при несбалансированности классов. Таким образом, если в выборке 95 отрицательных и 5 положительных объектов, то при  точности 0.95 мы можем не угадать не один положительный класс.
Это означает, что доля правильных ответов сама по себе не несет никакой информации о качестве работы алгоритма a(x), и вместе с ней следует анализировать соотношение классов в выборке.
 Также полезно вместе с долей правильных ответов вычислять долю правильных ответов алгоритма в каждом классе.

## 8. Матрица ошибок

Выше мы убедились, что в случае с несбалансированными классами одной доли правильных ответов недостаточно  необходимы еще метрики качества.

Но введем сначала понятие матрицы ошибок. Это способ разбить объекты на четыре категории в зависимости от комбинации истинного ответа и ответа алгоритма

## 9. Метрики

Гораздо более информативными критериями являются точность (precision) и полнота (recall). Точность показывает, какая доля объектов, выделенных классификатором как положительные, действительно является положительными. Полнота показывает, какая часть положительных объектов была выделена классификатором.
Рассмотрим, например, задачу предсказания реакции клиента банка на звонок
с предложением кредита. Ответ y = 1 означает, что клиент возьмет кредит после рекламного звонка, ответ y = −1  что не возьмет.
Соответственно, планируется обзванивать только тех клиентов, для которых классификатор a(x) вернет ответ 1.
Если классификатор имеет высокую точность, то практически все клиенты, которым будет сделано предложение, откликнутся на него.
Если классификатор имеет высокую полноту, то предложение будет сделано практически всем клиентам, которые готовы откликнуться на него.

Отметим, что точность и полнота не зависят от соотношения размеров классов. Даже если объектов положительного класса на порядки меньше, чем объектов отрицательного класса, данные показатели будут корректно отражать качество работы алгоритма.

Существует несколько способов получить один критерий качества на основе точности и полноты. Один из них  F-мера, гармоническое среднее точности и полноты: