1 Azure ML Studio

1.1 Данные

Исходные данные[1] – 25000 изображений собак и кошек.

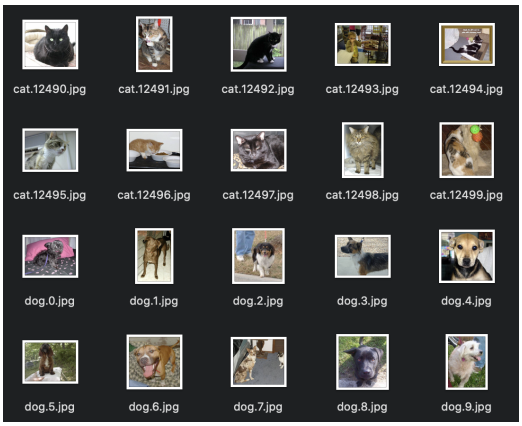

Рис. 1: Исходные данные.

Для работы с изображениями, найдены их гистограммы – характеристики распределения интенсивности изображения. Данные представлены в формате CSV и содержат: название файла (`FileName`), 512 столбцов числовых значений от 0 до 1 (`X1` ... `X512`) и отклик (`Label`).

Для обучения воспользуемся лишь частью данных и загрузим набор данных `CATS_DOGS.csv` в Azure ML. Разделим данные на тренировочную и тестовую части, в отношении 75/25. Для этого зададим параметр `Fraction of rows in the first output dataset` равным `0.75`. Тогда на первом выходе данные для обучения модели, второй выход – для ее оценки. Значение `Random seed` укажем для одинаковых результатов случайного разделения данных.

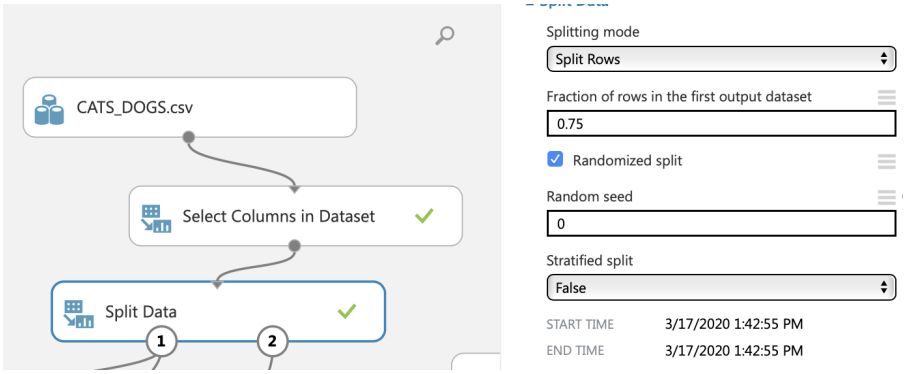

Рис. 2: Разделение данных.

[1] https://www.kaggle.com/c/dogs-vs-cats

1.2 Модель классификатора

Для обучения модели SVM классификатора используется блок `Two-Class Support Vector Machine` из раздела `Machine Learning`. Два интересующих нас параметра – `Random number seed` и `Lambda`. Остальные параметры остаются заданными по умолчанию.

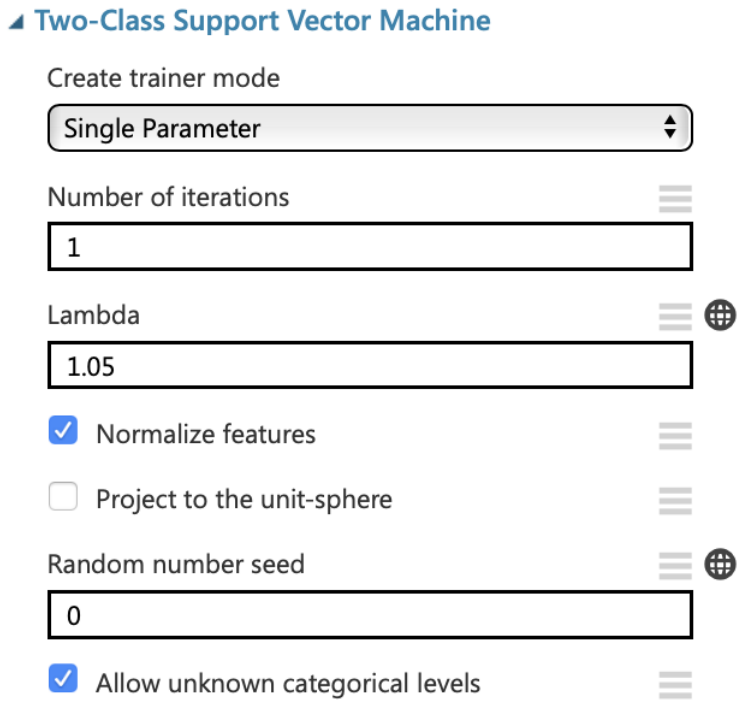

Рис. 3: Параметры блока Two-Class Support Vector Machine.

Блок `Train Model` все также отвечает за обучение модели. На вход подаются данные и выбранный метод машинного обучения. В качестве данных выступают колонки `X1` ... `X512` и отклик `Label`. В параметрах данного блока необходимо выбрать столбец данных, соответствующий отклику.

После запуска модели, полученные значения коэффициентов уравнения гиперплоскости можно посмотреть в параметрах блока `Train Model`, пункт `Visualize`. Параметр `Bias` (смещение) соответствует коэффициенту θ_0, а названия столбцов данных соответствующим коэффициентам $\theta_1, \ldots, \theta_p$.

1.3 Задача классификации

Для решения задачи классификации необходимы данные. В качестве данных могут выступать либо данные в формате CSV, либо это могут быть введенные вручную значения с помощью блока `Enter Data Manually`. В рассматриваемой задаче – это набор данных `CATS_DOGS_FOR_PREDICTION.csv` из которого нужны лишь определенные строки. Например с помощью блока `Apply SQL Transformation` и запроса вида:

```
select * from t1 where FileName in ('cat.1006.jpg', 'dog.1046.jpg');
```

выберем строки данных для изображений `cat.1006.jpg` и `dog.1046.jpg`.

Затем с помощью `Select Columns in Dataset` выберем только предикторы `X1` ... `X512`. После запуска эксперимента `Run` результаты классификации доступны в пункте `Visualize` блока `Score Model`.

К набору данных добавляются колонки `Scored Probabilities` и `Scored Labels`. В первой указана вероятность отнесения объекта к положительному классу, а во второй – результат бинарной классификации. Положительный класс назначается, если вероятность больше или равна 0.5. При большом количестве данных обязательно используйте фильтрацию данных, например выберете только колонку `Scored Labels` с помощью блока `Select Columns in Dataset`.

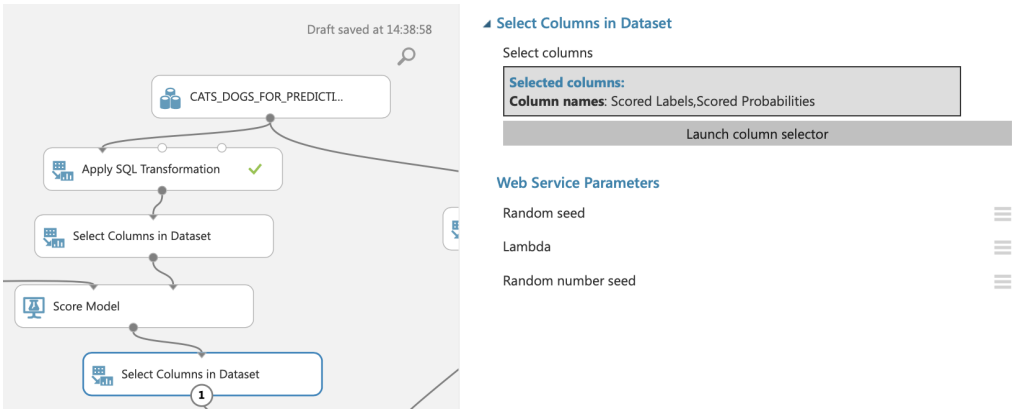

Рис. 4: Этапы подготовки данных.

2 ROC-анализ

За оценку модели отвечает блок `Evaluate Model`, подключаемый к `Score Model`, при этом, к блоку `Score Model` должны быть подключены тестовые данные, содержащие все предикторы и отклик – то есть те 25 процентов от исходных данных. После обучения модели и запуска доступны:

- `Confusion matrix` (матрица ошибок):

Матрица ошибок		Верный класс	
		+	−
Прогноз	+	TP	FP
	−	FN	TN

- `Precision` (точность) – это доля объектов, действительно являющихся положительными к тем, что названы положительными в результате классификации:
$$\texttt{Precision} = \frac{\texttt{TP}}{\texttt{TP} + \texttt{FP}}.$$
- `Recall` (полнота) – это доля объектов, классифицированных, как положительные, к тем, что действительно являются положительными. Также называется долей истинно положительных примеров `TPR` (True Positives Rate):
$$\texttt{Recall} = \frac{\texttt{TP}}{\texttt{TP} + \texttt{FN}}.$$
- `AUC` (площадь под кривой).

Ползунок для значения `Threshold` позволяет изменять порог отсечения, тем самым влияя на результат бинарной классификации.

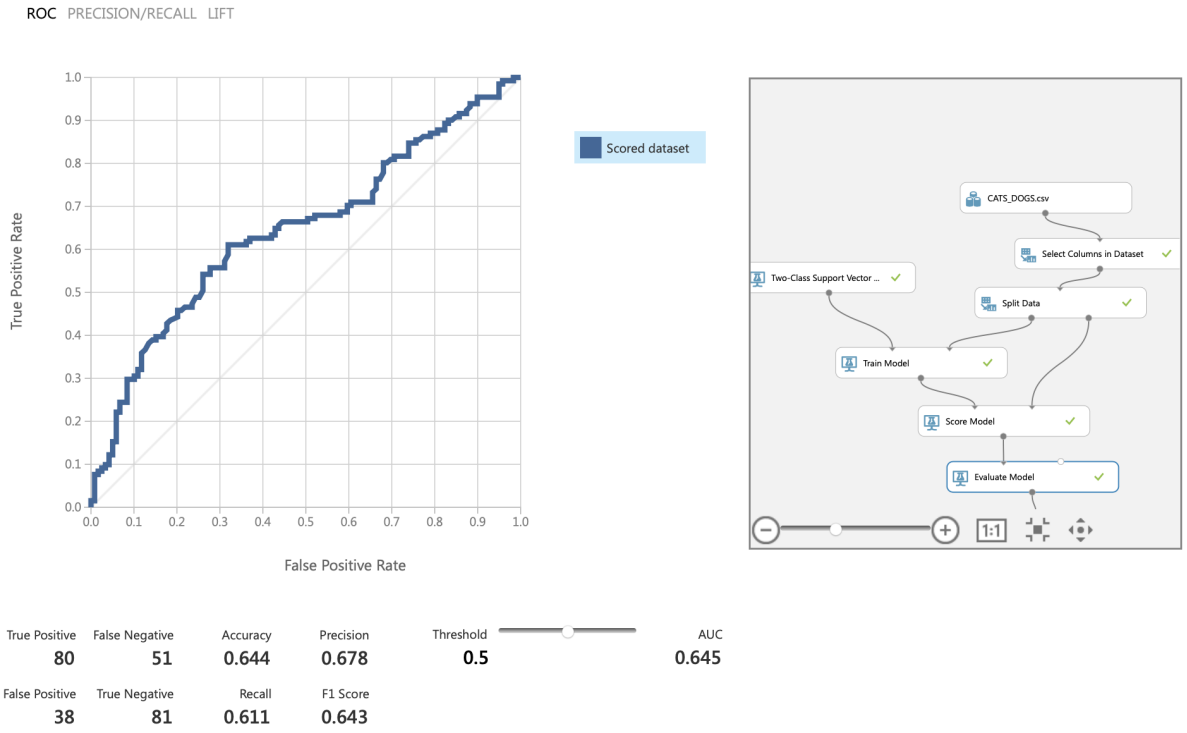

Рис. 5: Оценка модели и значения метрик.