Google 机器学习课程笔记

1. 使用 tf.estimator API

构建基本框架：

```
import tensorflow as tf

# Set up a linear classifier.
classifier = tf.estimator.LinearClassifier()

# Train the model on some example data.
classifier.train(input_fn=train_input_fn, steps=2000)

# Use it to predict.
predictions = classifier.predict(input_fn=predict_input_fn)
```

train_input_fn 是什么？

```
train_input_fn = tf.estimator.inputs.numpy_input_fn({"x": x_train},
y_train, batch_size=1, num_epochs=None, shuffle=False)
```

predict_input_fn 是什么？

```
predict_input_fn = tf.estimator.inputs.numpy_input_fn({"x": samples},
batch_size=1, num_epochs=1, shuffle=False)
```

2. pandas 学习
   https://colab.research.google.com/notebooks/mlcc/intro_to_pandas.ipynb?hl=zh-cn
3. TensorFlow 基本步骤

https://colab.research.google.com/notebooks/mlcc/first_steps_with_tensor_flow.ipynb?hl=zh-cn

4. 合成特征和离群值

https://colab.research.google.com/notebooks/mlcc/synthetic_features_and_outliers.ipynb?hl=zh-cn

5.数据集小时用交叉训练
大时用 10-15%做测试

6.validation
在每次迭代时，我们都会对训练数据进行训练并评估测试数据，并以基于测试数据的评估结果为指导来选择和更改各种模型超参数，这种方法的问题是
多次重复执行该流程可能导致我们不知不觉地拟合我们的特定测试集的特性。

7.

l1 正则化
起到稀疏的效果

• l2 正则化

根据奥卡姆剃刀定律，或许我们可以通过降低复杂模型的复杂度来防止过拟合，这种原则称为正则化。

也就是说，并非只是以最小化损失（经验风险最小化）为目标：

minimize(Loss(Data|Model))

而是以最小化损失和复杂度为目标，这称为结构风险最小化：

minimize(Loss(Data|Model) + complexity(Model))

模型开发者通过以下方式来调整正则化项的整体影响：用正则化项的值乘以名为 lambda（又称为正则化率）的标量。也就是说，模型开发者会执行以下运算：

$$\text{minimize}(\text{Loss}(\text{Data}|\text{Model})+\lambda \ \text{complexity}(\text{Model}))$$

执行 L_2 正则化对模型具有以下影响

- 使权重值接近于 0（但并非正好为 0）
- 使权重的平均值接近于 0，且呈正态（钟形曲线或高斯曲线）分布。

在选择 lambda 值时，目标是在简单化和训练数据拟合之间达到适当的平衡：

- 如果您的 lambda 值过高，则模型会非常简单，但是您将面临数据欠拟合的风险。您的模型将无法从训练数据中获得足够的信息来做出有用的预测。
- 如果您的 lambda 值过低，则模型会比较复杂，并且您将面临数据过拟合的风险。您的模型将因获得过多训练数据特点方面的信息而无法泛化到新数据。

逻辑回归中的正则化

大多数逻辑回归模型会使用以下两个策略之一来降低模型复杂性：

- L_2 正则化。
- 早停法，即，限制训练步数或学习速率。
- L1 正则化

8.离线训练和在线训练
离线训练适用于变化不大的（如大型图片集）

9.标签泄露，使模型带有欺骗性
如癌症医院案例

10.特征工程

（1）

映射字符串值：转为 onehot

映射分类（枚举）值

eg：分类特征具有一组离散的可能值。例如，名为 Lowland Countries 的特征只包含 3 个可能值：

```
{'Netherlands', 'Belgium', 'Luxembourg'}
```

可以表示为 3 个单独的布尔值特征：

- x_1：是荷兰吗？
- x_2：是比利时吗？
- x_3：是卢森堡吗？

采用这种方法编码还可以简化某个值可能属于多个分类这种情况（例如，“与法国接壤”对于比利时和卢森堡来说都是 True）

（2）数据清理
标准化（两种方法：均匀分布的标准化，正态分布的标准化）

处理极端离群值方法：取对数，最大值“限制”为某个任意值（在最大点存在一个小峰值）

分箱

11. 特征组合 (Feature Crosses)：对非线性规律进行编码

在实践中，机器学习模型很少会组合连续特征。不过，机器学习模型却经常组合独热特征矢量，将独热特征矢量的特征组合视为逻辑连接

特征组合的种类

我们可以创建很多不同种类的特征组合。例如：

- [A X B]：将两个特征的值相乘形成的特征组合。
- [A x B x C x D x E]：将五个特征的值相乘形成的特征组合。
- [A x A]：对单个特征的值求平方形成的特征组合。

Eg：预测加州房价的一个好的特征组合：[binned latitude X binned longitude X binned roomsPerPerson]

将分箱纬度与分箱经度组合可以让模型了解 roomsPerPerson 特定于城市的效果。分箱可防止纬度变化与经度变化产生相同的效果。根据箱的精细程度，此特征组合可以反映出特定于城市、特定于社区，甚至特定于街区的效果。