**特征组合 (Feature Crosses)**

**1.对非线性规律进行编码**

在图 1 和图 2 中，我们做出如下假设：

* 蓝点代表生病的树。
* 橙点代表健康的树。



图 1. 这是线性问题吗？

您可以画一条线将生病的树与健康的树清晰地分开吗？当然可以。这是个线性问题。这条线并不完美。有一两棵生病的树可能位于“健康”一侧，但您画的这条线可以很好地做出预测。

现在，我们来看看下图：



图 2. 这是线性问题吗？

您可以画一条直线将生病的树与健康的树清晰地分开吗？不，您做不到。这是个非线性问题。您画的任何一条线都不能很好地预测树的健康状况。

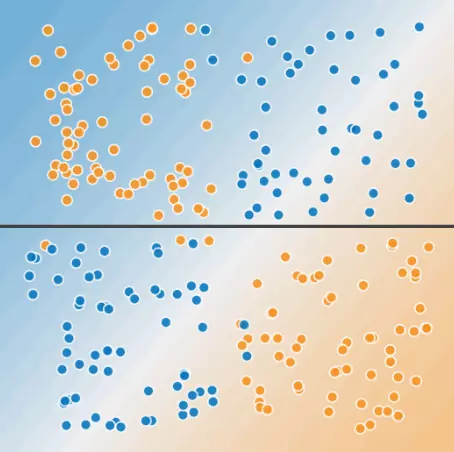


图 3. 一条线无法分开两类数据

要解决图 2 所示的非线性问题，可以创建一个特征组合。**特征组合**是指通过将两个或多个输入特征相乘来对特征空间中的非线性规律进行编码的合成特征。“cross”（组合）这一术语来自 [*cross product*](https://wikipedia.org/wiki/Cross_product)（向量积）。我们通过将 x\_{1}与x\_{2}组合来创建一个名为x\_{3}的特征组合：x\_{3}=x\_{1}x\_{2}

我们像处理任何其他特征一样来处理这个新建的x\_{3} 特征组合。线性公式变为：  
y=b+w\_{1}x\_{1}+w\_{2}x\_{2}+w\_{3}x\_{3}

线性算法可以算出w\_{3}的权重，就像算出w\_{1}和w\_{2}的权重一样。换言之，虽然w\_{3}表示非线性信息，但您不需要改变线性模型的训练方式来确定w\_{3}的值。

**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。

通过采用[随机梯度下降法](https://developers.google.com/machine-learning/crash-course/reducing-loss/stochastic-gradient-descent)，可以有效地训练线性模型。因此，在使用扩展的线性模型时辅以特征组合一直都是训练大规模数据集的有效方法。

**关键字词**

1. 特征组合 (feature cross)

通过将单独的特征进行组合（相乘或求笛卡尔积）而形成的[**合成特征**](https://developers.google.com/machine-learning/crash-course/glossary#synthetic_feature)。特征组合有助于表示非线性关系。

1. 合成特征 (synthetic feature)

一种[**特征**](https://developers.google.com/machine-learning/crash-course/glossary#feature)，不在输入特征之列，而是从一个或多个输入特征衍生而来。合成特征包括以下类型：

* 将一个特征与其本身或其他特征相乘（称为[**特征组合**](https://developers.google.com/machine-learning/crash-course/glossary#feature_cross)）。
* 两个特征相除。
* 对连续特征进行[**分桶**](https://developers.google.com/machine-learning/crash-course/glossary#bucketing)，以分为多个区间分箱。

通过[**标准化**](https://developers.google.com/machine-learning/crash-course/glossary#normalization)或[**缩放**](https://developers.google.com/machine-learning/crash-course/glossary#scaling)单独创建的特征不属于合成特征。

**2.组合独热矢量**

到目前为止，我们已经重点介绍了如何对两个单独的浮点特征进行特征组合。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。例如，假设我们具有以下两个特征：国家/地区和语言。对每个特征进行独热编码会生成具有二元特征的矢量，这些二元特征可解读为 country=USA, country=France 或 language=English, language=Spanish。然后，如果您对这些独热编码进行特征组合，则会得到可解读为逻辑连接的二元特征，如下所示：

country:usa AND language:spanish

再举一个例子，假设您对纬度和经度进行分箱，获得单独的独热 5 元素特征矢量。例如，指定的纬度和经度可以表示如下：

binned\_latitude = [0, 0, 0, 1, 0]

binned\_longitude = [0, 1, 0, 0, 0]

假设您对这两个特征矢量创建了特征组合：

binned\_latitude X binned\_longitude

此特征组合是一个 25 元素独热矢量（24 个 0 和 1 个 1）。该组合中的单个 1 表示纬度与经度的特定连接。然后，您的模型就可以了解到有关这种连接的特定关联性。

假设我们更粗略地对纬度和经度进行分箱，如下所示：

binned\_latitude(lat) = [

0 < lat <= 10

10 < lat <= 20

20 < lat <= 30

]

binned\_longitude(lon) = [

0 < lon <= 15

15 < lon <= 30

]

针对这些粗略分箱创建特征组合会生成具有以下含义的合成特征：

binned\_latitude\_X\_longitude(lat, lon) = [

0 < lat <= 10 AND 0 < lon <= 15

0 < lat <= 10 AND 15 < lon <= 30

10 < lat <= 20 AND 0 < lon <= 15

10 < lat <= 20 AND 15 < lon <= 30

20 < lat <= 30 AND 0 < lon <= 15

20 < lat <= 30 AND 15 < lon <= 30

]

现在，假设我们的模型需要根据以下两个特征来预测狗主人对狗狗的满意程度：

* 行为类型（吠叫、叫、偎依等）
* 时段  
  如果我们根据这两个特征构建以下特征组合：

[behavior type X time of day]

我们最终获得的预测能力将远远超过任一特征单独的预测能力。例如，如果狗狗在下午 5 点主人下班回来时（快乐地）叫喊，可能表示对主人满意度的正面预测结果。如果狗狗在凌晨 3 点主人熟睡时（也许痛苦地）哀叫，可能表示对主人满意度的强烈负面预测结果。

线性学习器可以很好地扩展到大量数据。对大规模数据集使用特征组合是学习高度复杂模型的一种有效策略。[神经网络](https://developers.google.com/machine-learning/crash-course/introduction-to-neural-networks)可提供另一种策略。

**关键字词**

1. 独热编码又称为one-hot 编码 (one-hot encoding)

一种稀疏向量，其中：

* 一个元素设为 1。
* 所有其他元素均设为 0。

one-hot 编码常用于表示拥有有限个可能值的字符串或标识符。例如，假设某个指定的植物学数据集记录了 15000 个不同的物种，其中每个物种都用独一无二的字符串标识符来表示。在特征工程过程中，您可能需要将这些字符串标识符编码为 one-hot 向量，向量的大小为 15000。

**3.Playground 练习**

**特征组合简介**

特征组合真的能使模型拟合非线性数据吗？要找出答案，请尝试完成此练习。

**任务**：尝试创建一个能够通过手动更改以下三个输入特征的权重将蓝点与橙点分开的模型：

* x\_{1}
* x\_{2}
* x\_{1}x\_{2}（特征组合）

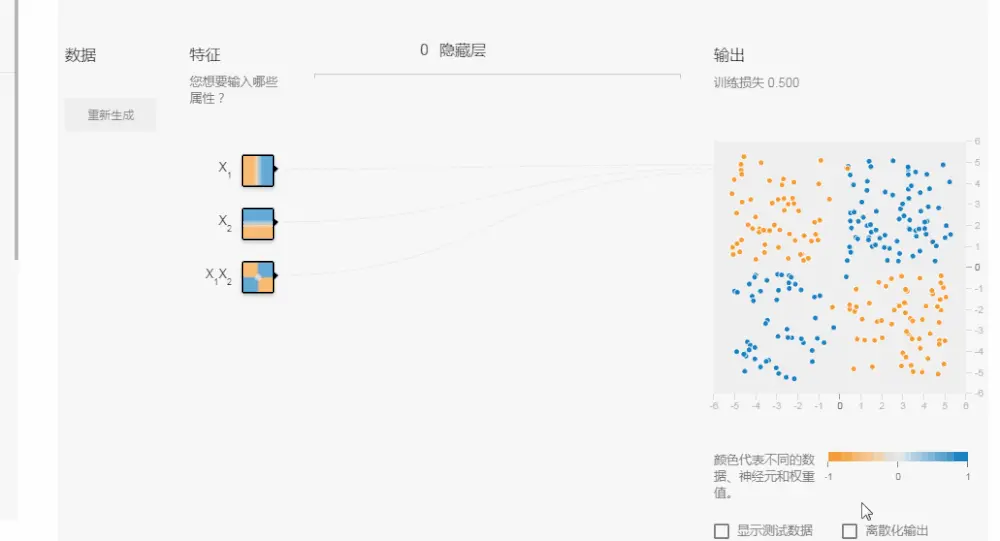
要手动更改权重，请执行以下操作：

1. 点击一条将“特征”连接到“输出”的线。随后将出现一个输入表单。
2. 在该输入表单中输入一个浮点值。
3. 按 Enter 键。

请注意，此练习的界面不包含“Step”（步）按钮。这是因为此练习不会迭代地训练模型。您将手动输入模型的“最终”权重。  
答案如下:  
w\_{1} = 0  
w\_{2} = 0  
x\_{1}x\_{2} = 1（或任何正值）



如果您为特征组合输入一个负值，则该模型会将蓝点与橙点分开，但预测结果是完全错误的。也就是说，该模型会将蓝点预测为橙点，并将橙点预测为蓝点。

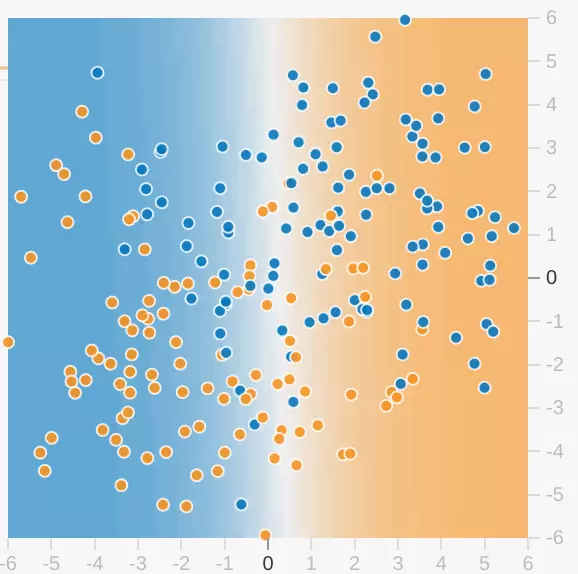


**更复杂的特征组合**

现在，我们来使用一些高级特征组合。此 Playground 练习中的数据集看起来有点像飞镖游戏中的飞镖击中位置，中间是蓝点，外圈是橙点。

关于模型直观图示的说明,具体如下:

每个 Playground 练习都会显示模型当前状态的直观图示。例如，以下就是一个模型的直观图示：  
请注意以下关于模型直观图示的说明：



* 每个蓝点表示一类数据的一个样本（例如，一棵健康的树）。
* 每个橙点表示另一类数据的一个样本（例如，一棵生病的树）。
* 背景颜色表示该模型对于应该在何处找到相应颜色样本的预测。某个蓝点周围显示蓝色背景表示该模型正确地预测了该样本。相反，某个蓝点周围显示橙色背景则表示该模型错误地预测了该样本。
* 背景的蓝色和橙色部分色调会有深浅之分。例如，直观图示的左侧是纯蓝色，但在直观图示的中心颜色则逐渐淡化为白色。您可以将颜色强度视为表明该模型对其猜测结果的自信程度。因此，纯蓝色表示该模型对其猜测结果非常自信，而浅蓝色则表示该模型的自信程度稍低。（图中所示的模型直观图示在预测方面的表现非常糟糕。）

可以通过直观图示来判断模型的进展。（“非常棒 - 大多数蓝点都有蓝色背景”或者“糟糕！蓝点有橙色背景。”）除了颜色之外，Playground 还会以数字形式显示模型的当前损失。（“糟糕！损失正在上升，而不是下降。”）

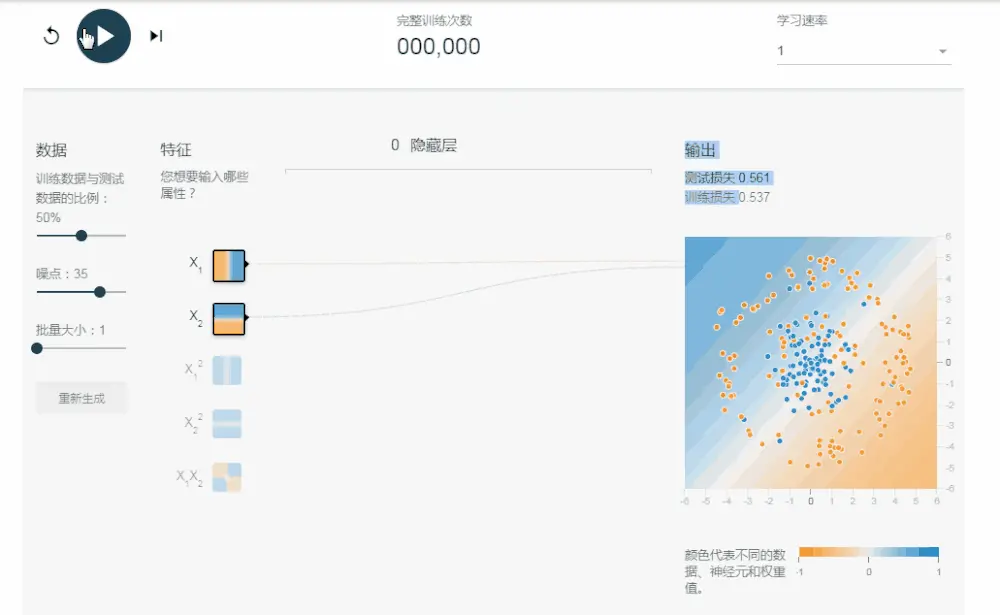
**任务 1：**

按照指定方式运行此线性模型。花一两分钟时间（不再更多）尝试不同的学习速率设置，看看是否出现任何改进。线性模型可以针对此数据集得出有效结果吗？  
特征x\_{1},x\_{2},学习速率设置为3

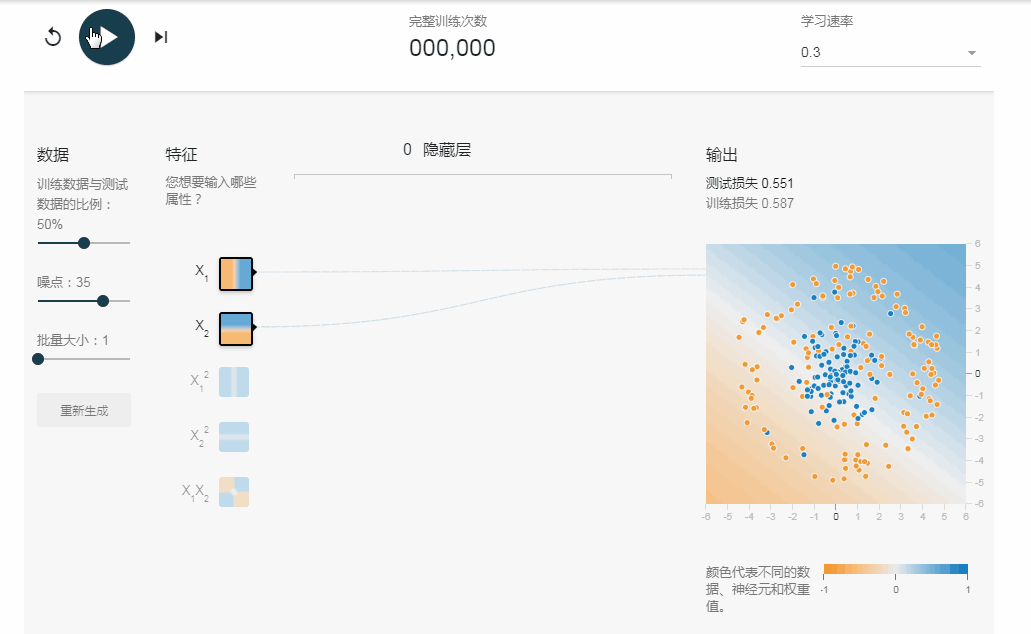
输出  
测试损失 0.960  
训练损失 0.840



特征x\_{1},x\_{2},学习速率设置为1  
输出  
测试损失 0.972  
训练损失 0.688



特征x\_{1},x\_{2},学习速率设置为0.3  
输出  
测试损失 0.820  
训练损失 0.657



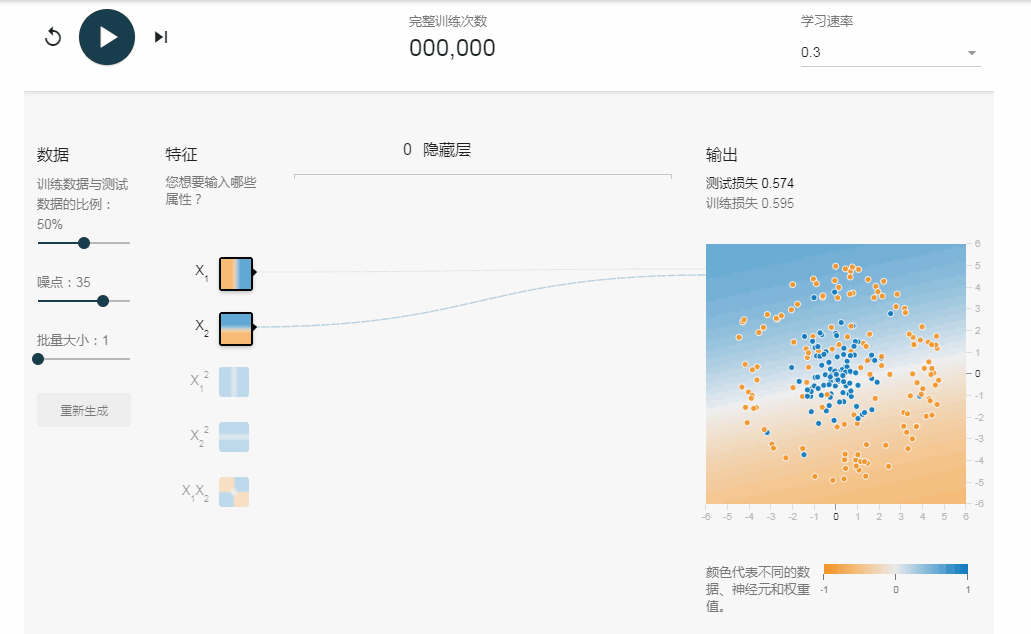
总结:不可以。线性模型无法有效地为此数据集建模。降低学习速率可以减少损失，但损失仍然会收敛于不可接受的高值。

**任务 2：**

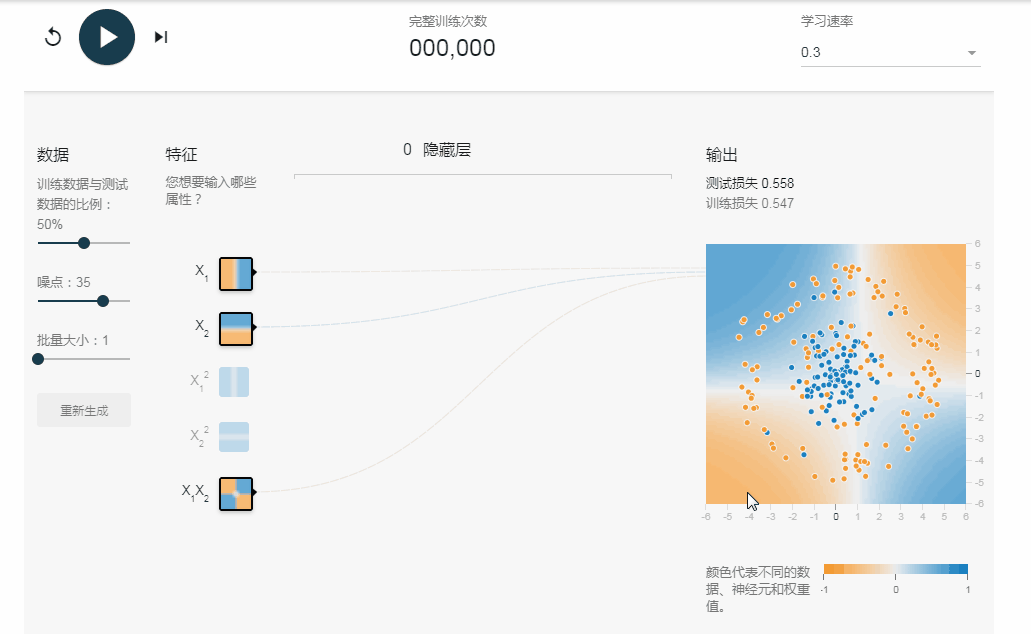
现在，尝试添加向量积特征（例如x\_{1}x\_{2} ），以尝试优化效果。

* 哪些特征最有帮助？
* 您可以实现的最佳效果是什么？

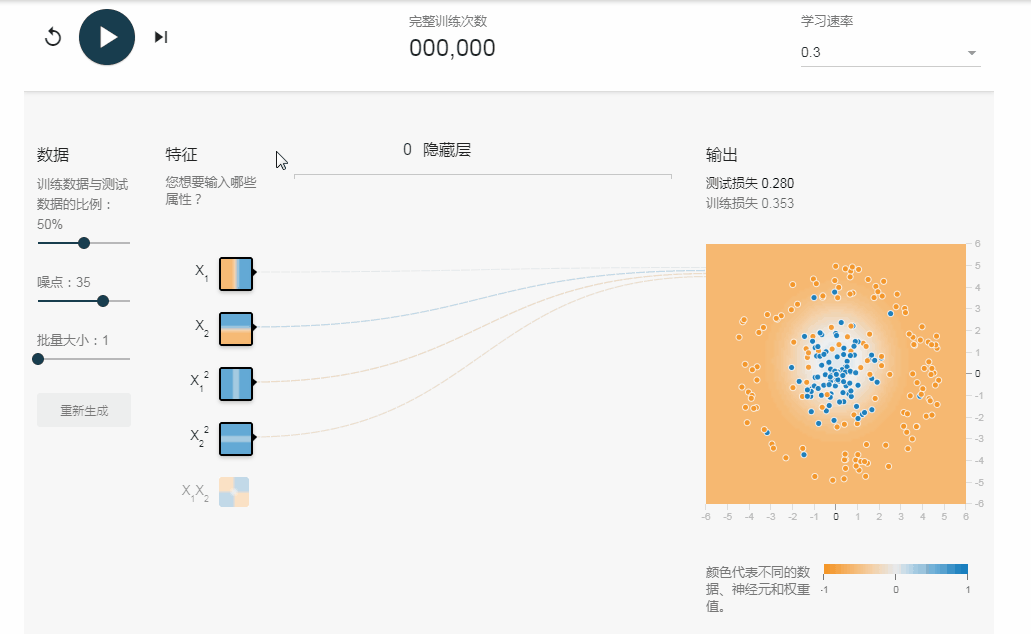
特征x\_{1},x\_{2},学习速率设置为0.3  
输出结果  
测试损失 0.908  
训练损失 0.658



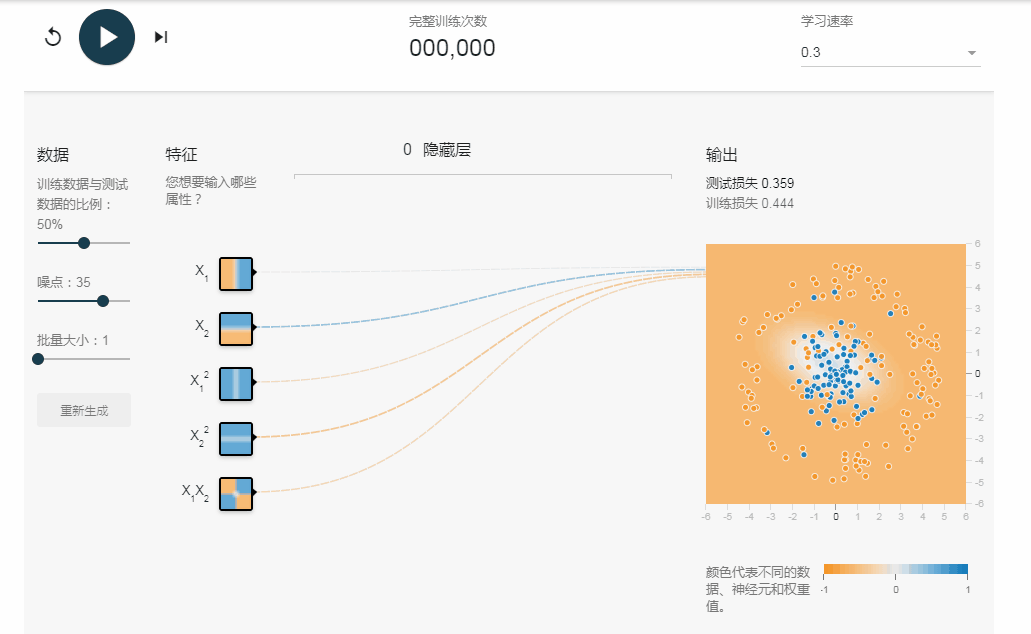
特征x\_{1},x\_{2},x\_{1}x\_{2},学习速率设置为0.3  
输出结果  
测试损失 0.717  
训练损失 0.577



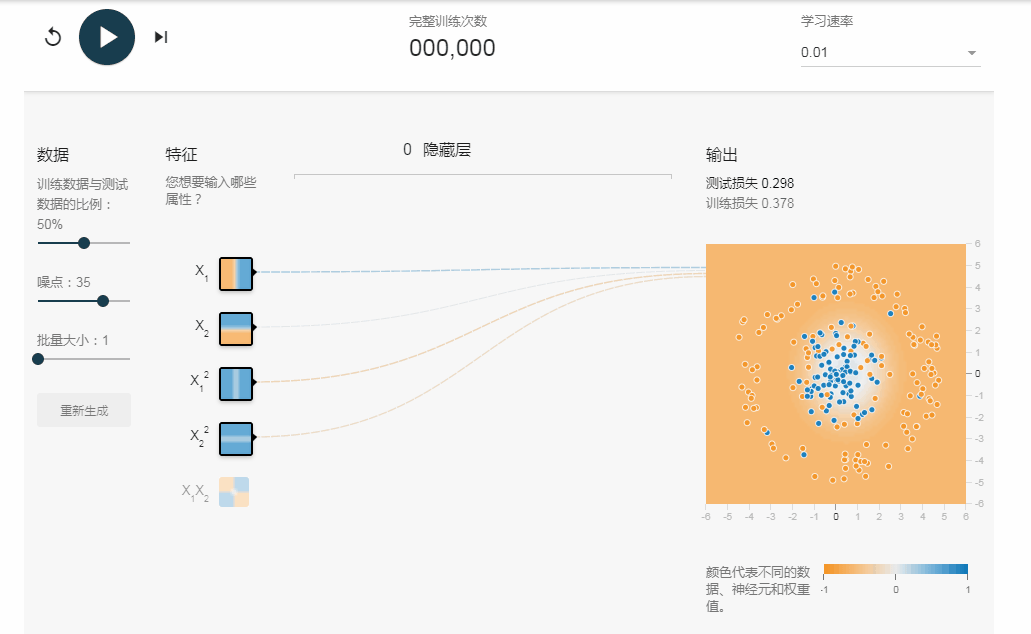
特征x\_{1}^{2} ,x\_{2}^{2},x\_{1},x\_{2},学习速率设置为0.3  
输出  
测试损失 0.245  
训练损失 0.320



特征x\_{1}^{2} ,x\_{2}^{2},x\_{1},x\_{2},x\_{1}x\_{2},学习速率设置为0.3  
输出  
测试损失 0.232  
训练损失 0.309



特征x\_{1}^{2} ,x\_{2}^{2},x\_{1},x\_{2},x\_{1}x\_{2},学习速率设置为0.01  
输出  
测试损失 0.172  
训练损失 0.208



总结:

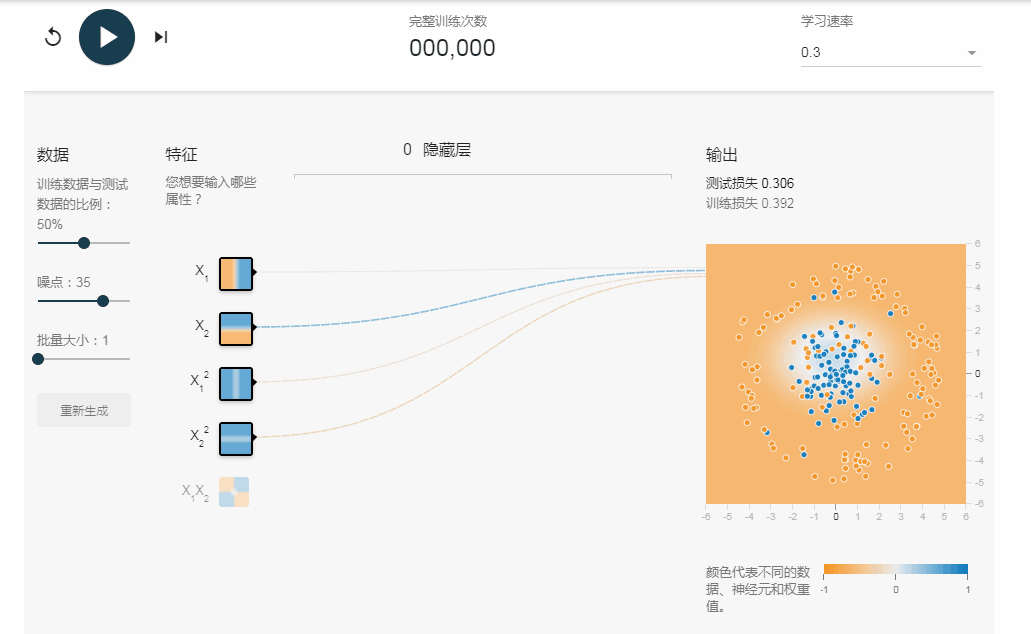
* x\_{1}^{2} ,x\_{2}^{2},x\_{1},x\_{2}这些特征有帮助,此外当特征组合有有x\_{1}^{2} ,x\_{2}^{2},另外添加将 x\_{1}x\_{2}特征组合似乎没什么帮助。
* 降低学习速率,对测试损失以及训练损失确实能够降低.

备注:Playground 的数据集是随机生成的。因此，每个人的答案不一定完全一致。事实上，如果您每次运行都重新生成数据集，那么您自己的结果与您之前运行的结果也不一定完全一致。

**任务 3：**

如果您有一个出色的模型，请检查该模型的输出面（以背景颜色显示）。

1. 看起来像线性模型吗？
2. 您会如何描述该模型？



该模型输出面看起来不像线性模型，反而像椭圆形。

**4.检测练习**

加利福尼亚州不同城市的[房价](https://developers.google.com/machine-learning/crash-course/california-housing-data-description)有很大差异。假设您必须创建一个模型来预测房价。以下哪组特征或特征组合可以反映出特定城市中 roomsPerPerson 与房价之间的关系？  
A、三个独立的分箱特征：[binned latitude]、[binned longitude]、[binned roomsPerPerson]  
B、一个特征组合：[binned latitude X binned longitude X binned roomsPerPerson]  
C、两个特征组合：[binned latitude X binned roomsPerPerson] 和 [binned longitude X binned roomsPerPerson]  
D、一个特征组合：[latitude X longitude X roomsPerPerson]

答案是B,具体解析如下:

A、三个独立的分箱特征：[binned latitude]、[binned longitude]、[binned roomsPerPerson]  
建议您采用分箱方式，因为这样可以让模型了解单个特征内的非线性关系。不过，一个城市存在于多个维度上；因此，要了解特定于城市的关系，需要对纬度与经度进行组合。

B、一个特征组合：[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。

C、两个特征组合：[binned latitude X binned roomsPerPerson] 和 [binned longitude X binned roomsPerPerson]  
建议您采用分箱方式；不过，城市是纬度和经度的结合体，因此单独的特征组合会导致模型无法了解特定于城市的价格。

D、一个特征组合：[latitude X longitude X roomsPerPerson]  
在本例中，不建议您组合实值特征。例如，将纬度的实值与 roomsPerPerson 组合后，一个特征（比如纬度）上发生 10% 的变化就相当于另一个特征（比如 roomsPerPerson）上发生 10% 的变化。

作者：durian221783310  
链接：https://www.jianshu.com/p/e6d6f6a991d0  
来源：简书  
简书著作权归作者所有，任何形式的转载都请联系作者获得授权并注明出处。