3.4.3. Использование композиций алгоритмов для отбора признаков

Сами по себе решающие деревья не очень полезны, но они очень активно используются при построении композиций, в частности, в случайных лесах и в градиентном бустинге над деревьями. В данных композициях измерить важность признака можно аналогичным образом: суммируется уменьшение критерия информативности R_j по всем деревьям композиции, и чем больше данная сумма, тем важнее признак j для композиции. То есть признаки оцениваются с помощью того, насколько сильно они смогли уменьшить значение критерия информативности в совокупности по всем деревьям композиции.

Для случайного леса можно предложить еще один интересный способ оценивания информативности признаков. В этой композиции каждое базовое дерево b_n обучается по подмножеству объектов обучающей выборки. Таким образом, есть объекты, на которых дерево не обучалось, и набор этих объектов является валидационной выборкой для дерева n. Такая выборка называется out-of-bag. Итак, метод заключается в следующем: ошибку Q_n базового дерева b_n оценивают по out-of-bag-выборке и запоминают. После этого признак j превращают в абсолютно бесполезный, шумовой: в матрицу «объекты-признаки» все значения в столбце j перемешивают. Затем то же самое дерево b_n применяют к данной выборке с перемешанным признаком j и оценивают качество дерева на out-of-bag-подвыборке. Q'_n — ошибка out-of-bag-подвыборке, она будет тем больше, чем сильнее дерево использует признак j. Если он активно используется в дереве, то ошибка сильно уменьшится, поскольку значение данного признака испорчено. Если же данный признак совершенно не важен для дерева и не используется в нем, то ошибка практически не изменится. Таким образом, информативность признака j оценивают как разность

$$Q'_n - Q_n.$$

Далее эти информативности усредняют по всем деревьям случайного леса, и чем больше будет среднее значение, тем важнее признак. На практике оказывается, что информативности, вычисленные описанным образом, и информативности, вычисленные как сумма уменьшения критерия информативности, оказываются очень связаны между собой.

3.5. Понижение размерности

3.5.1. Примеры использования методов понижения размерности

Зачем нужно понижать размерность и чем этот подход отличается от отбора признаков? Для ответа на этот вопрос полезно рассмотреть несколько примеров.

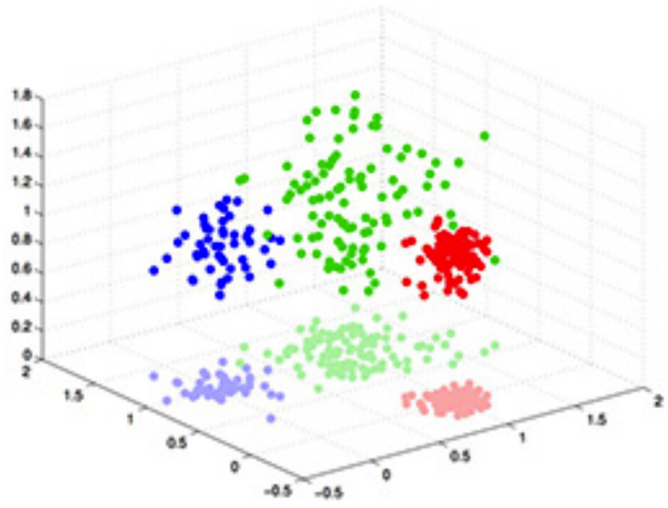

Рис. 3.3: Пример выборки, для которой необходимо решить задачу классификации

На рисунке 3.3 изображена выборка с тремя размерностями. Видно, что если убрать из нее признак, отложенный по оси Z, получится двумерная выборка, в которой синий, зеленый и красный кластеры будут разделимы даже линейными методами.

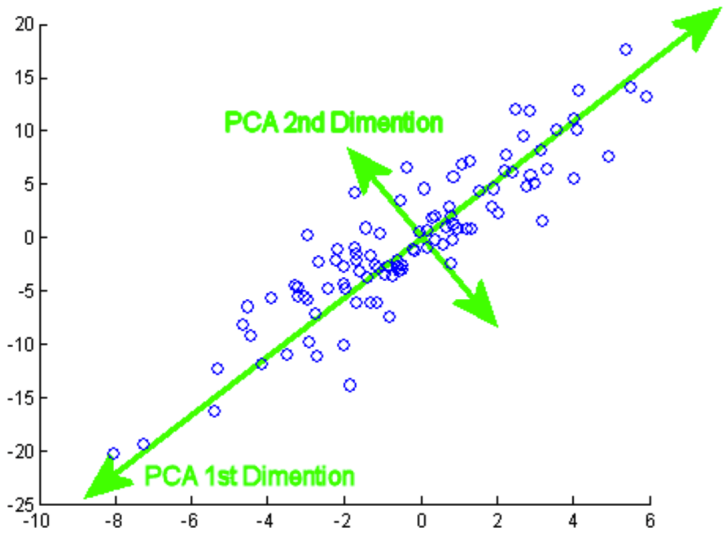

Рис. 3.4: Пример выборки с линейно зависимыми признаками

На рисунке 3.4 показан более сложный случай. В данной выборке оба признака значимые, но при этом они линейно зависимые, и этим можно воспользоваться, чтобы устранить избыточность в данных. Однако отбора признаков для этого не хватит: необходимо сформировать новый признак на основе двух исходных.

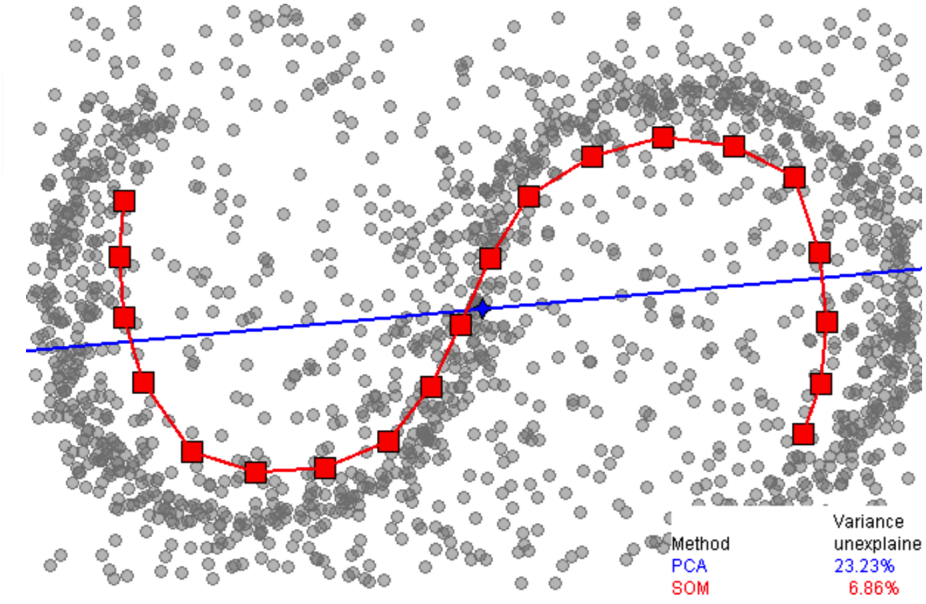

Рис. 3.5: Пример данных, которые необходимо спроецировать на кривую

Бывают и еще более сложные случаи, как, например, на рисунке 3.5. Здесь также можно спроецировать выборку на некоторую кривую, но при этом кривая очень нелинейная, и её, скорее всего, будет сложно найти.

Эти примеры подводят к задаче понижения размерности, которая состоит в формировании новых признаков на основе исходных. При этом количество признаков становится меньше, но они должны сохранять в себе как можно больше информации, присутствующей в исходных признаках.

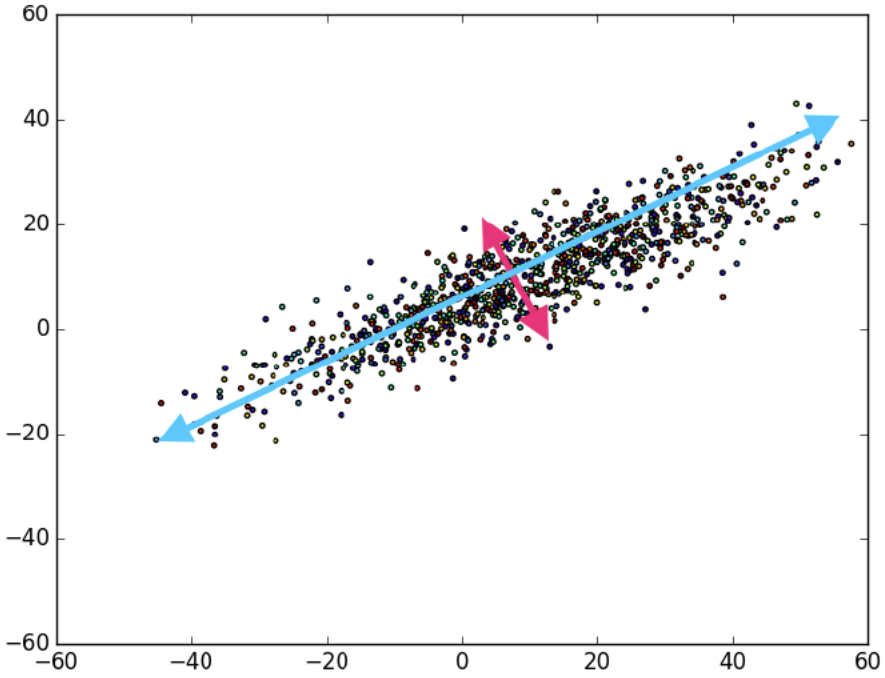

Рис. 3.7: Пример выборки, для которой необходимо решить задачу понижения размерности

3.7. Метод главных компонент: решение

3.7.1. Вывод решения задачи метода главных компонент

Ранее была описана формулировка задачи метода главных компонент, теперь необходимо её решить. Одна из постановок задачи метода главных компонент — это максимизация дисперсии:

$$\begin{cases} \sum\limits_{j=1}^{d} \mathrm{w}_j^T X^T X \mathrm{w}_j \to \max\limits_{W} \\ W^T W = I \end{cases}$$

В первой строке записана дисперсия после проецирования, а во второй — ограничение, обеспечивающее наличие единственного решения.

В методе главных компонент есть один нюанс: выражение, через которое записана дисперсия, будет означать именно дисперсию выборки только в том случае, если матрица объекты-признаки центрирована (среднее каждого признака равно нулю). Далее считается, что, выборка центрирована, и среднее из каждого столбца в матрице объекты-признаки уже вычли.

Итак, чтобы разобраться, как устроено решение этой задачи, необходимо сначала рассмотреть простой частный случай: требуется найти ровно одну компоненту, на которую проецируется вся выборка, так, чтобы дисперсия после проецирования была максимальной:

$$\begin{cases} \mathrm{w}_1^T X^T X \mathrm{w}_1 \to \max\limits_{\mathrm{w}_1} \\ \mathrm{w}_1^T \mathrm{w}_1 = 1 \end{cases}$$

Для решения подобных задач условной оптимизации необходимо выписать лагранжиан:

$$L(\underset{1}{\mathrm{w}}, \lambda) = \overset{T}{\underset{1}{\mathrm{w}}} X^T X \underset{1}{\mathrm{w}} - \lambda(\overset{T}{\underset{1}{\mathrm{w}}} \underset{1}{\mathrm{w}} - 1).$$