

10 главных алгоритмов машинного обучения

Методы машинного обучения можно разделить на 3 основные категории: контролируемое, неконтролируемое и подкрепляемое обучение.

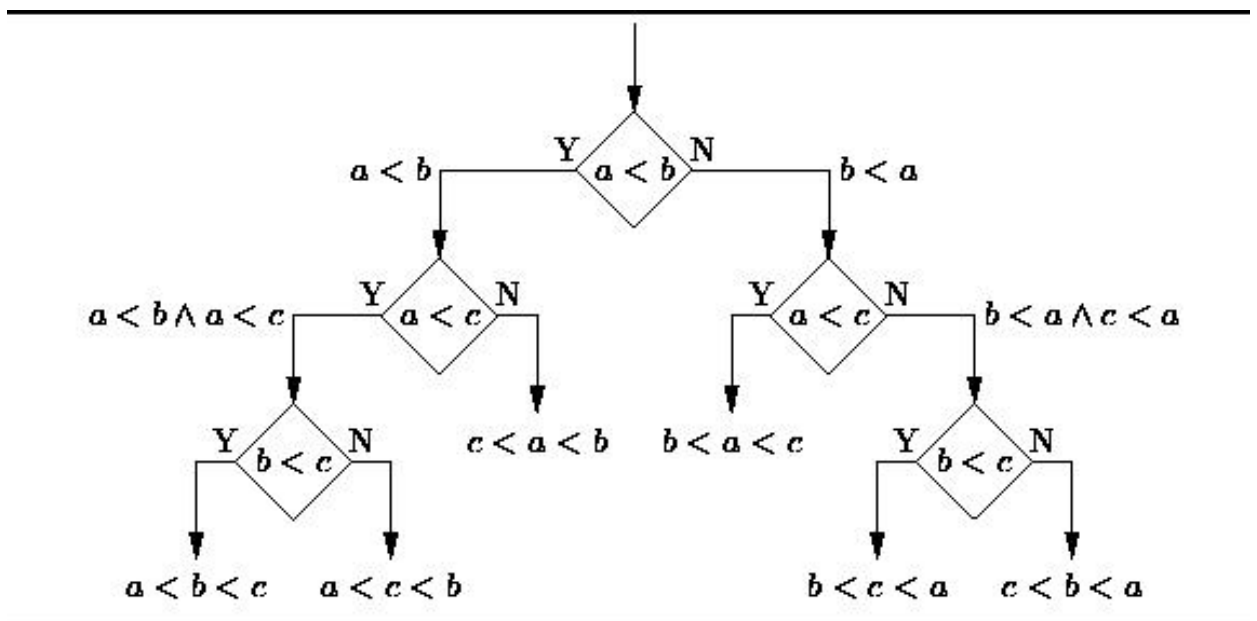
Контролируемое обучение полезно в тех случаях, когда свойство (ярлык) доступно для определенного массива данных (обучающего набора), но на данный момент оно отсутствует и должно быть предсказано для других случаев. Неконтролируемое обучение используется для обнаружения неявных отношений в данном немаркированном наборе данных.

Подкрепляемое обучение — что-то среднее между вышеописанными категориями: есть некоторая форма обратной связи, доступная для каждого шага или действия, но отсутствует ярлык и сообщение об ошибке. Так как это были вводные занятия, я не узнал ничего о подкрепляемой категории, но надеюсь, что эти 10 алгоритмов, касающиеся контролируемого и неконтролируемого обучения, вас заинтересуют.

Контролируемое обучение

Дерево принятия решений

[Дерево принятия решений](#) — средство поддержки принятия решений, которое использует древовидный граф или модель принятия решений, а также возможные последствия их работы, включая вероятность наступления события, затраты ресурсов и полезность. На рисунке 1 подано графическое представление структуры дерева.

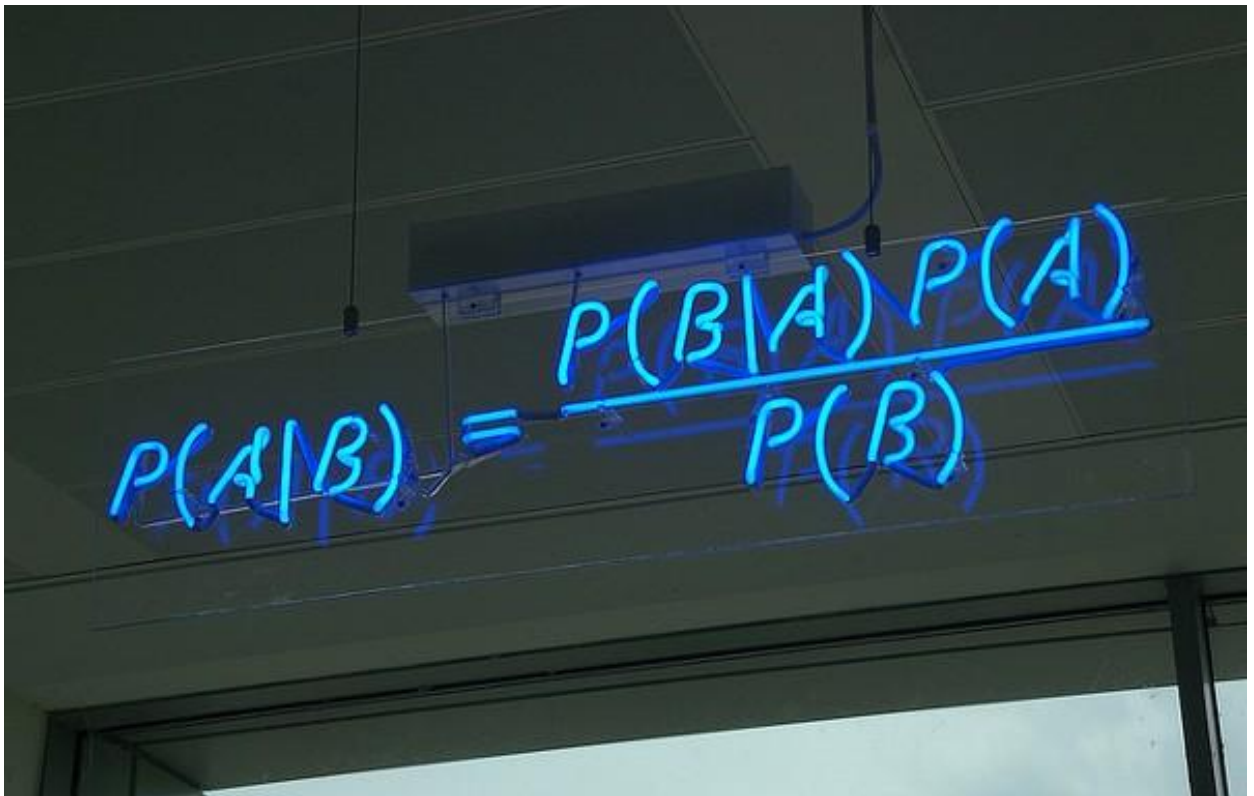


С точки зрения бизнес-решения, дерево классификации является минимальным количеством вопросов «да/нет», ответив на которые, можно

сделать верный выбор. Если рассматривать дерево как метод, то оно позволяет подойти к решению проблемы со структурированной и систематической стороны, чтобы в итоге прийти к логическому выводу.

Наивная байесовская классификация

Наивные байесовские классификаторы представляют собой семейство простых вероятностных классификаторов, которые основаны на применении Теоремы Байеса со строгими (наивными) предположениями о независимости функций. На приведенном ниже изображении указано равенство; здесь $P(A|B)$ является вероятностью гипотезы A при наступлении события B (апостериорная вероятность), $P(B|A)$ — вероятностью наступления события B при истинности гипотезы A , $P(A)$ — априорной вероятностью гипотезы A и $P(B)$ — полной вероятностью наступления события B .

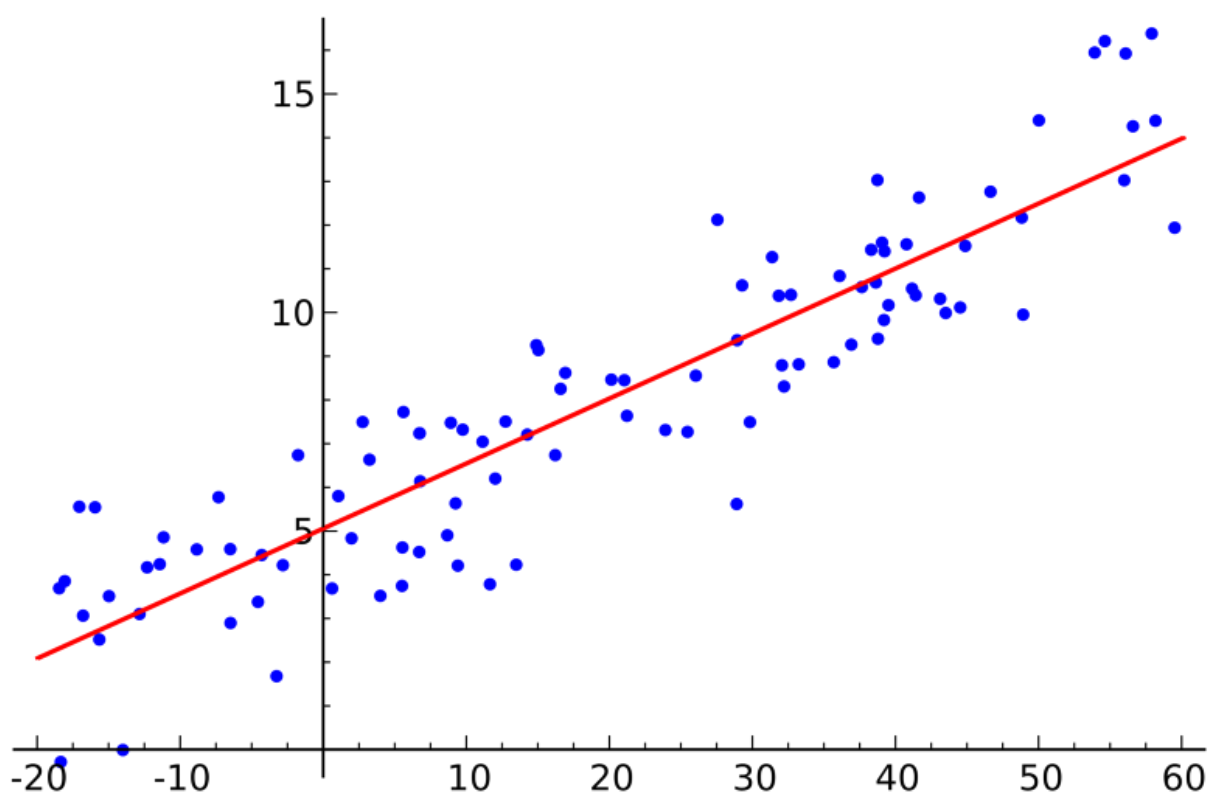

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

Абстрагируясь от теории и переходя к практике, можно выделить следующие сферы применения Теоремы Байеса:

- «отлов» спама в электронной почте;
- сегментация новостных статей по их тематике;
- определение эмоционального окраса блока текста;
- программное обеспечение для распознавания лиц.

Метод наименьших квадратов

Если вы знакомы со статистикой, то наверняка слышали о [линейной регрессии](#) ранее. Наименьшие квадраты выступают в роли метода для реализации линейной регрессии. Чаще всего она представляется в виде задачи подгонки прямой линии, проходящей через множество точек. Есть несколько вариантов ее осуществления, и метод наименьших квадратов — один из них. Можно нарисовать линию, а затем измерить расстояние по вертикали от каждой точки к линии и «перенести» эту сумму вверх. Необходимой линией будет та конструкция, где сумма расстояний будет минимальной. Иными словами, кривая проводится через точки, имеющие нормально распределенное отклонение от истинного значения.

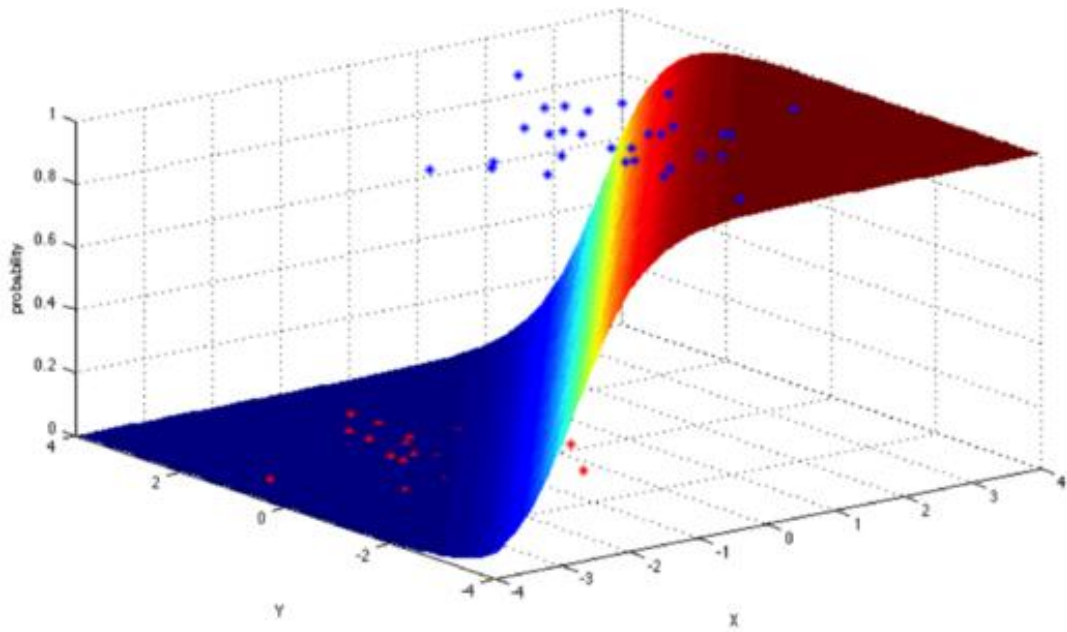


Если линейная функция применима для подбора данных, то метод наименьших квадратов относится к типам метрики ошибок, которая минимизирует погрешности.

Логистическая регрессия

[Логистическая регрессия](#) представляет собой мощный статистический способ прогнозирования вероятности возникновения некоторого события с одной или несколькими независимыми переменными. Логистическая регрессия определяет степень зависимости между категориальной зависимой и одной или несколькими независимыми переменными путем использования

логистической функции, являющейся аккумулятивным логистическим распределением.

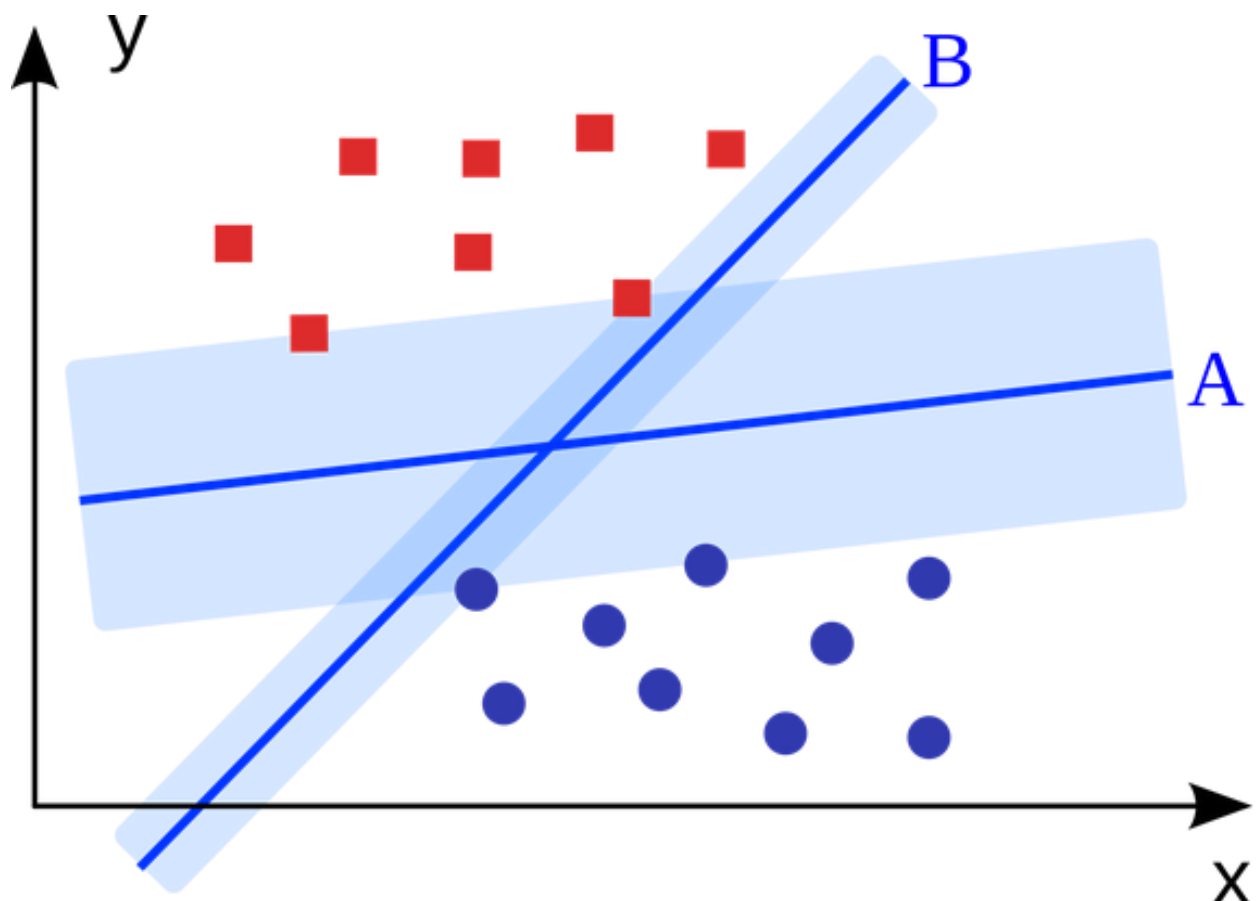


Данный алгоритм активно используется в реальной жизни, а именно при:

- оценке кредитоспособности лица (кредитном скоринге);
- измерении показателей успешности маркетинговых кампаний;
- предсказании доходов с определенного продукта;
- вычислении возможности возникновения землетрясения в конкретный день.

Метод опорных векторов

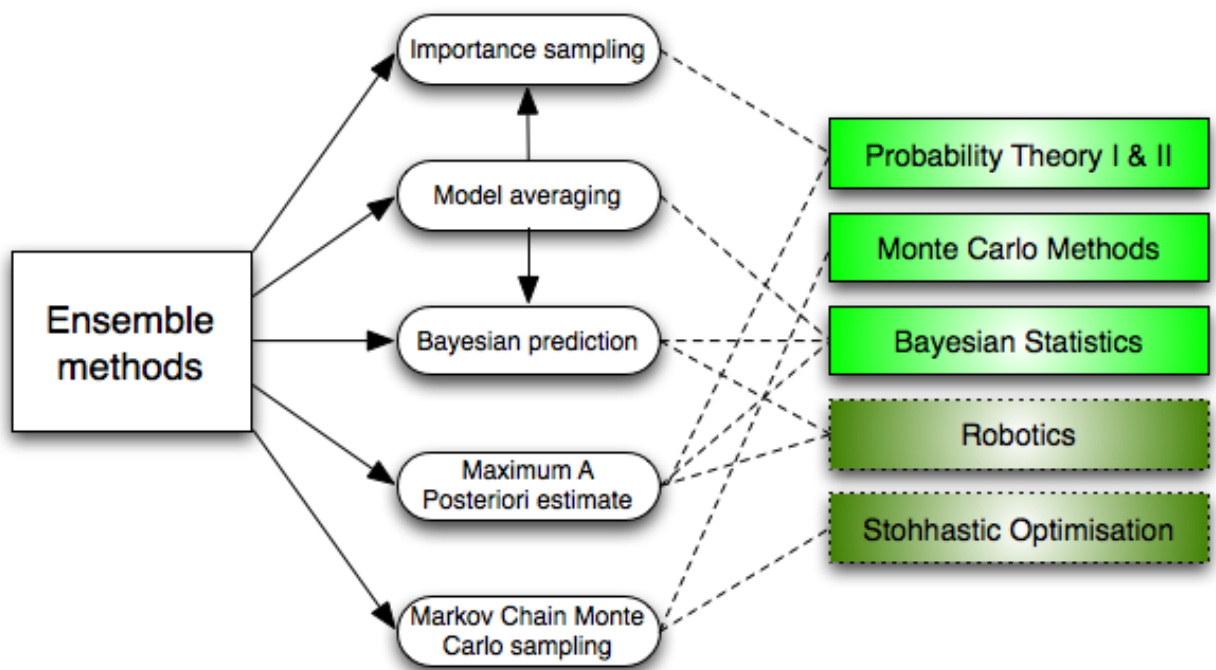
Метод опорных векторов (SVM) — это набор алгоритмов, использующихся для задач классификации и регрессионного анализа. Учитывая, что в N -мерном пространстве каждый объект принадлежит одному из двух классов, SVM генерирует $(N-1)$ -мерную гиперплоскость с целью разделения этих точек на 2 группы. Это как если бы вы на бумаге изобразили точки двух разных типов, которые можно линейно разделить. Помимо того, что метод выполняет сепарацию объектов, SVM подбирает гиперплоскость так, чтобы та характеризовалась максимальным удалением от ближайшего элемента каждой из групп.



Среди наиболее масштабных проблем, которые были решены с помощью метода опорных объектов (и его модифицированных реализаций) выделяют отображение рекламных баннеров на сайтах, распознавание пола на основании фотографии и [сплайсинг человеческой ДНК](#).

Метод ансамблей

[Метод ансамблей](#) основан на обучающих алгоритмах, которые формируют множество классификаторов, а затем сегментируют новые точки данных, отталкиваясь от голосования или усреднения. Оригинальный метод ансамблей — не что иное, как Байесовское усреднение, но более поздние алгоритмы включают исправления ошибок выходного кодирования, бэггинг (bagging) и бустинг (boosting). Бустинг направлен на превращение слабых моделей в сильные путем построения ансамбля классификаторов. Бэггинг также агрегирует усовершенствованные классификаторы, но используется при этом параллельное обучение базовых классификаторов. Говоря языком математической логики, бэггинг — улучшающее объединение, а бустинг — улучшающее пересечение.



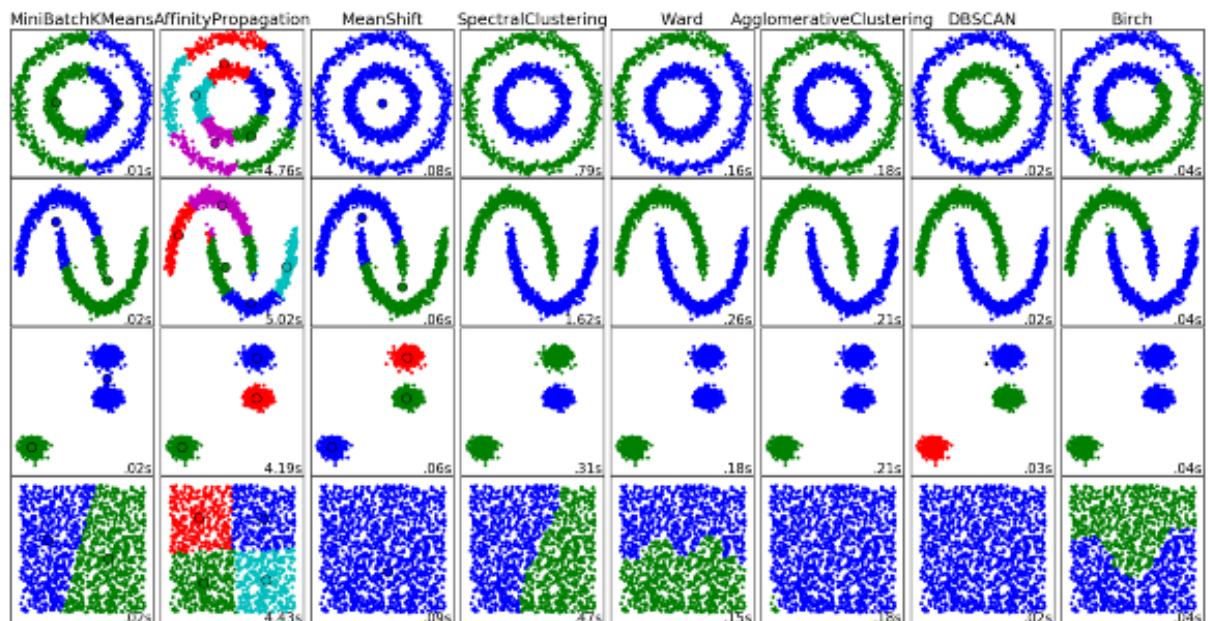
Все же, почему метод ансамблей превосходит отдельно стоящие прогнозные модели?

- **Он минимизирует влияние случайностей.** Агрегированный классификатор «усредняет» ошибку каждого из базовых классификаторов — соответственно, влияние случайностей на усредненную гипотезу существенно уменьшается.
- **Он снижает дисперсию.** Совокупное мнение целого множества моделей лучше, чем мнение отдельно взятой модели. В экономике это называется диверсификацией — расширение ассортимента выпускаемой продукции повышает эффективность производства и предотвращает банкротство. Ансамбль моделей имеет больший шанс найти глобальный оптимум, поскольку поиск идет из разных точек исходного множества гипотез.
- **Он предотвращает выход за пределы множества.** Вероятен следующий случай: агрегированная гипотеза находится за пределами множества базовых гипотез. При построении комбинированной гипотезы любым путем (логистическая регрессия, усредненное значение, голосование), множество гипотез расширяется, следовательно, полученный результат не выходит за его рамки.

Неконтролируемое обучение

Алгоритмы кластеризации

Задача кластеризации состоит в группировании множества объектов таким образом, чтобы поместить максимально похожие между собой элементы в одну группу (кластер).



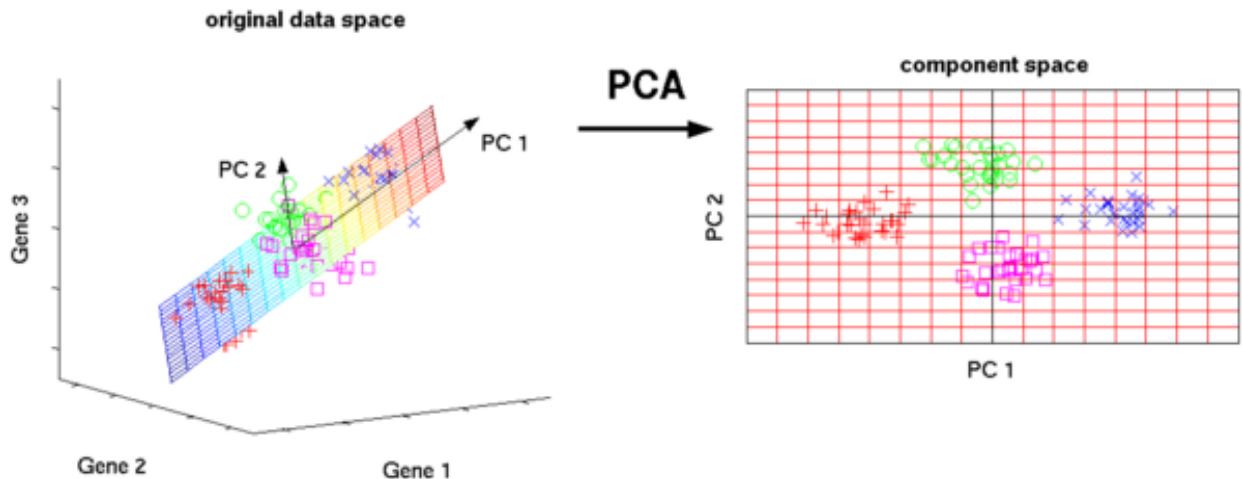
Алгоритмов кластеризации существует довольно много, и все они отличаются друг от друга. Самые популярные из них:

- алгоритмы на базе центра тяжести треугольника;
- алгоритмы на основе подключения;
- алгоритмы плотности на основе пространственной кластеризации;
- вероятностный алгоритм;
- алгоритм уменьшения размерности;
- нейронные сети и машинное обучение.

Алгоритмы кластеризации используются в биологии, социологии и информационных технологиях. Например, в биоинформатике с помощью кластеризации анализируются сложные сети взаимодействующих генов, состоящие порой из сотен или даже тысяч элементов. А при анализе результатов социологических исследований рекомендуется осуществлять анализ методом Уорда, при котором внутри кластеров оптимизируется минимальная дисперсия, в итоге создаются группы приблизительно равных размеров.

Метод главных компонент

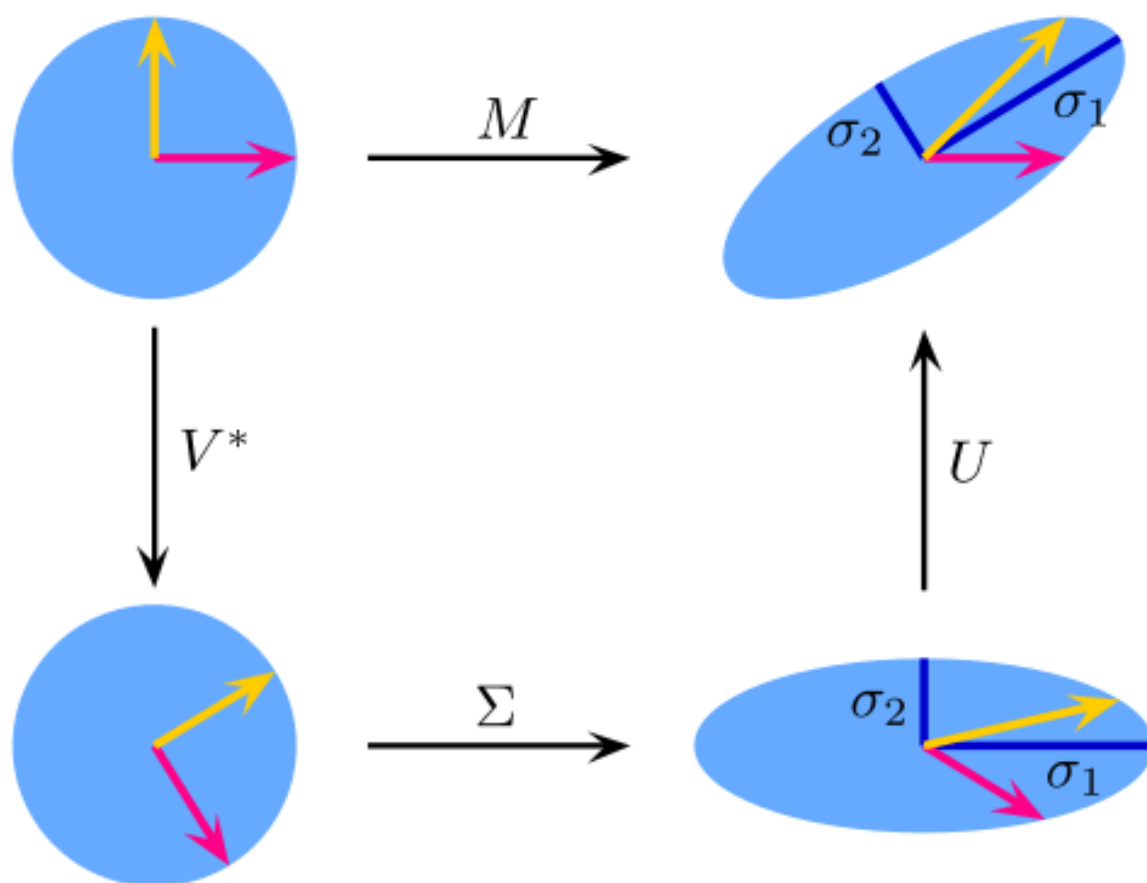
Метод главных компонент (PCA) — это статистическая процедура, которая использует ортогональное преобразование с целью конвертации набора наблюдений за возможно коррелированными переменными в набор значений линейно некоррелированных переменных, называемых главными компонентами.



Отдельные области применения PCA включают в себя сжатие и упрощение данных для облегчения обучения, а также визуализацию. Решение об использовании метода главных компонент зависит от уровня познания предметной области. PCA не подходит для применения в случаях с плохо упорядоченными данными (все компоненты метода имеют довольно высокую дисперсию).

Сингулярное разложение

В линейной алгебре под сингулярным разложением (SVD) понимают разложение прямоугольной вещественной или комплексной матрицы. Для матрицы M размерностью $[m \times n]$ существует такое разложение, что $M = U\Sigma V$, где U и V — унитарные матрицы, а Σ - диагональная матрица.



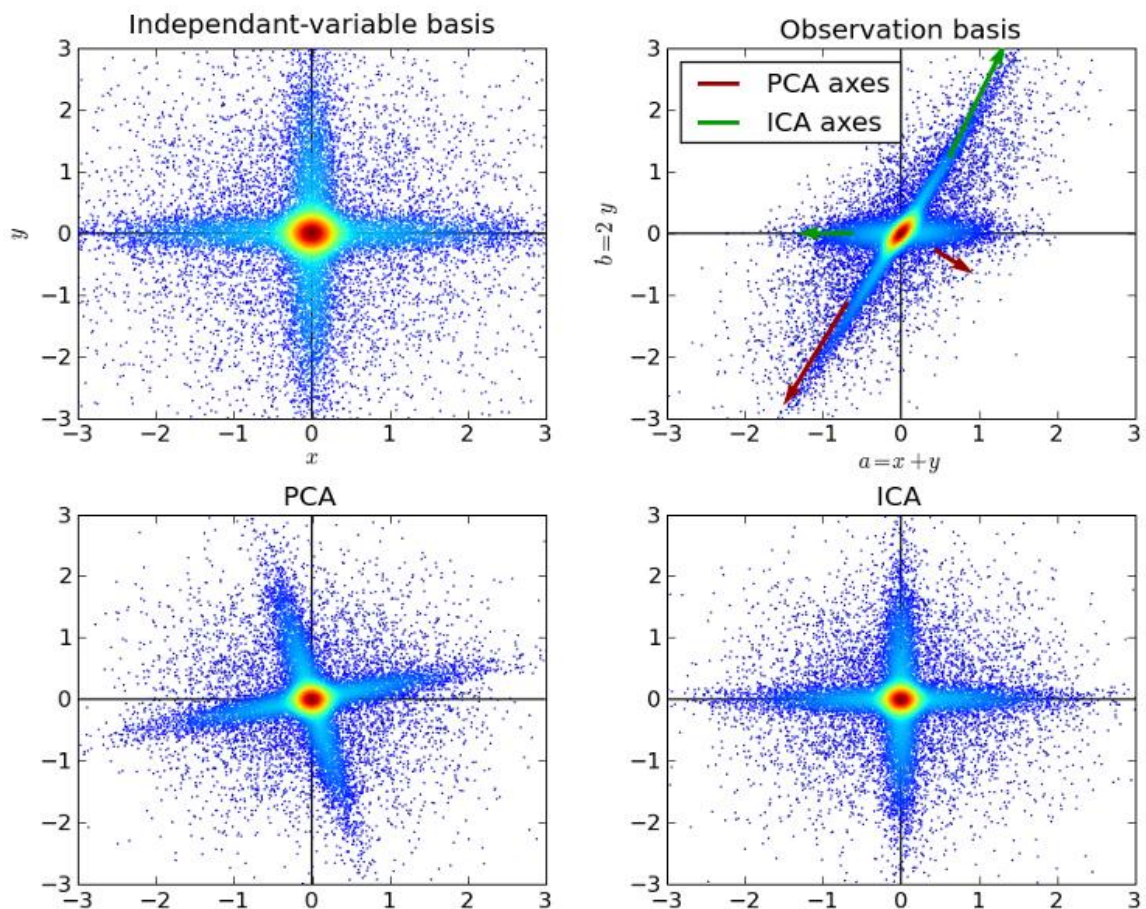
$$M = U \cdot \Sigma \cdot V^*$$

Метод главных компонент является простым применением сингулярного разложения. Первые алгоритмы компьютерного видения использовали PCA и SVD, чтобы представить [лица в виде суммы базисных компонент](#), выполнить уменьшение размерности, а затем сопоставить их с изображениями из обучающей выборки. И хотя современные методы характеризуются более сложной реализацией, многие из них по-прежнему работают на базе подобных алгоритмов.

Анализ независимых компонент

[Анализ независимых компонент \(ICA\)](#) представляет собой статистический метод выявления скрытых факторов, которые лежат в основе множества случайных величин, сигналов и прочих измерений. ICA определяет порождающую модель для исследуемых многофакторных данных, которые обычно подаются в виде большой базы данных образцов. В модели переменные подаются как линейная смесь некоторых скрытых переменных, а любая информация о законах смешивания отсутствует. Предполагается, что

скрытые переменные независимы друг от друга и представляются как негауссовские сигналы, поэтому они называются независимыми компонентами исследуемых данных.



Анализ независимых компонент непосредственно связан с методом главных компонент, но это гораздо более мощная техника, способная найти скрытые факторы источников, когда классические методы в лице PCA дают сбой. Алгоритм ICA применяется в телекоммуникациях, астрономии, медицине, распознавании речи и изображений, диагностировании и тестировании сложных электронных систем и, наконец, поиске скрытых факторов и источников движения финансовых показателей.

Освоив каждый из вышеперечисленных алгоритмов машинного обучения, вы сможете создавать приложения на основе кластера технологий, которые с каждым годом продвигают науку и технику вперед.

Источник: <http://ru.datasides.com/code/algorithms-machine-learning/>