Условной плотностью называется величина $p(x|y) = \frac{p(x,y)}{p(y)} = \frac{p(y|x)p(x)}{p(y)} = \frac{p(y|x)p(x)}{\int p(y|x)p(x)dx}$ – теорема Байеса ($= \frac{\text{функция правдоподобия} * \text{априорное распределение}}{\text{нумеровочная константа}}$)

# Концепция машинного обучения

•Большинство математических задач сводятся к определению значений некоторых величин по заданному набору других величин

•Предполагается, что между величинами существует зависимость

•Выделяют два подхода к решению таких задач: математическое моделирование (model-based reasoning)и восстановление зависимостей путем обработки прошлого опыта

•Основное требование для применимости второго подхода – наличие обучающей информации

•Как правило в качестве таковой выступает выборка прецедентов– ситуационных примеров из прошлого

# Структура прецедента

•Атомарным элементом информации является прецедент — объект, обладающий набором наблюдаемых и скрытых переменных

•В обучающей выборке значения скрытых переменных известны

•В частности, особый интерес представляют ситуации, когда скрытые переменные объекта многомерные и связаны сложными взаимозависимостями

•Примерами таких объектов являются социальные сети, изображения, сигналы и др.

•Обучающая выборка при этом может состоять из одного сложного объекта!

•Требуется построить алгоритм, который позволял бы оценить зависимость между наблюдаемыми и скрытыми переменными объекта по обучающей выборке и использовать ее для обработки новых, не встречавшихся ранее объектов, значения скрытых переменных которых неизвестны

# Классификация

•Исторически возникла из задачи машинного зрения, поэтому часто употребляемый синоним –распознавание образов

•В классической задаче классификации обучающая выборка представляет собой набор отдельных объектов $(X, T) = \{(x_i, t_i)\}_{i=1}^{n}$

•У каждого объекта есть наблюдаемые переменные(признаки) $x_i = (x_{i,1}, \ldots, x_{i,d}) \in \mathbb{R}^d$ и скрытая переменная $t_i$, принимающая конечное число значений, обычно из множества $T = \{1, \ldots, l\}$

•Требуется построить алгоритм (классификатор), который по вектору признаковхвернул бы метку класса $\hat{t}$ или вектор оценок принадлежности (апостериорных вероятностей) к каждому из классов $\{p(t|x)\}_{t=1}^{l}$

# Регрессия

•Исторически возникла при исследовании влияния одной группы непрерывных случайных величин на другую группу непрерывных случайных величин

•В классической задаче восстановления регрессии обучающая выборка представляет собой набор отдельных объектов $(X, T) = \{(x_i, t_i)\}_{i=1}^{n}$

•У каждого объекта есть наблюдаемые переменные(признаки) $x_i = (x_{i,1}, \ldots, x_{i,d}) \in \mathbb{R}^d$ и скрытая переменная $t_i \in \mathbb{R}$

•Требуется построить алгоритм (регрессор), который по вектору признаков $x$ вернул бы точечную оценку значения регрессии $\hat{t}$, доверительный интервал $(t_-, t_+)$ или апостериорное распределение на множестве значений скрытой переменной $p(t|x)$

# Общая формулировка задач обработки данных

Все перечисленные выше задачи (и многие другие) обладают рядом общих черт

•Имеется массив объектов с наблюдаемыми переменными $X$ и скрытыми переменными $T$

•Предполагается, что между наблюдаемыми и скрытыми переменными существует зависимость

•Точный вид этой зависимости нам неизвестен и/или зависимость недетерминированная, т.е. значения наблюдаемых переменных не позволяют однозначно определить значения скрытых переменных

# Статистическая постановка задачи

•Каждый объект описывается парой $(x, t)$

•При статистической (вероятностной) постановке задачи машинного обучения предполагается, что обучающая выборка является набором независимых,

одинаково распределенных случайных величин, взятых из некоторой генеральной совокупности

При вероятностном подходе к решению этих задач, неопределенность в зависимости между $X$ и $T$ моделируется введением совместного распределения на все переменные $p(X, T)$. Выделяют два вида вероятностных моделей: порождающие (generative) и дискриминативные (discriminative)

# Порождающая модель

•При использовании порождающих моделей необходимо задать совместное распределение $p(X, T)$ на множестве объектов

•Зная совместное распределение, мы можем моделировать новые объекты из той же генеральной совокупности

•Если совместное распределение неизвестно (т.е. почти всегда), можно пробовать его настроить по обучающей выборке

•Как правило, это чрезвычайно сложная задача, решение которой, к тому же, избыточно в большинстве задач

# Дискриминативная модель

•При использовании дискриминативных моделей необходимо знать условное распределение $p(T|X)$ на множестве значений скрытых переменных объекта

•Зная условное распределение, мы можем определить наиболее вероятные значения скрытых переменных объекта

•В отличие от порождающей модели, дискриминативная модель не позволяет моделировать новые объекты из генеральной совокупности.

•В случае, когда условная плотность неизвестна, ее можно попробовать настроить по обучающей выборке

•Настройка дискриминативной модели более простая, поэтому если нам требуется только уметь определять значения скрытых переменных по наблюдаемым, использование такой модели предпочтительно

# Основные задачи, решаемые с помощью вероятностных моделей

•Обучение с учителем. Дана обучающая выборка $(X, T)$ и параметрически задано распределение $p(X, T|\theta)$. Задача: определить значение параметров $\theta$, при которых

распределение наилучшим образом описывает обучающую выборку $p(X,T|\theta) \rightarrow max_\theta$

•Обучение без учителя. Дана выборка объектов с наблюдаемыми переменными $X$ и параметрически задано распределение $p(X,T|\theta)$. Задача: определить значение параметров $\theta$, при которых распределение наилучшим образом описывает наблюдаемые данные $p(X|\theta) \rightarrow max_\theta$

•MAP-оценивание (метод оценки с помощью апостериорного максимума тесно связан с методом максимального правдоподобия (ML), но дополнительно при оптимизации использует априорное распределение величины, которую оценивает, Maximum a posteriori estimation). Дана вероятностная модель $p(X,T)$ и массив объектов с известными наблюдаемыми переменными $X$. Задача: определить наиболее вероятные значения скрытых переменных объектов $p(X,T) \rightarrow max_T$

•Оценка маргиналов. Дана вероятностная модель $p(X,T)$ и массив объектов с известными наблюдаемыми переменными $X$. Задача: определить маргинальное распределение на отдельно взятую скрытую переменную $p(t|X)$−?

Аналогичные задачи могут быть сформулированы и для дискриминативных моделей.

# Байесовские методы работы свероятностными моделями

•Для работы с вероятностными моделями обычно используются т.н. байесовские методы, базирующиеся на одноименном подходе к теории вероятностей:

| Подход | Частотный (классический) | Байесовский |
|---|---|---|
| Интерпретация случайности | Объективно неопределённое | Субъективное незнание |
| Величины | Случайные/неслучайные | Все величины случайные |
| Метод вывода (статистического оценивания) | Метод максимального правдоподобия (ML)<br>Проблема: $n \rightarrow \infty$ | Теорема Байеса |
| Оценки | Точечные (интервальные) | Апостериорное распределение |
| Применимость | $n \rightarrow \infty$ $(n \gg 1)$ | $\forall n$ |

•Основными достоинством байесовской парадигмы является возможность учета наших предпочтений на вид модели, характерные значения скрытых переменных, желаемые свойства оцениваемых параметров и пр.

•Как это ни удивительно, при решении практических задач таких предпочтений оказывается довольно много, хотя они не всегда лежат на поверхности

•Байесовский аппарат предоставляет удобное средство строгой формализации таких предпочтений

# Способы оценки обобщающей способности

•На сегодняшний день единственным универсальным способом оценивания обобщающей способности является кросс-валидация

• Байесовская регуляризация (МакКай, 1992: трактует параметры $\theta$ как случайные величины и добавляет их в вероятностную модель $p(X, T, \theta)$. Вопрос определения конкретного вида нашей вероятностной модели известен как задача выбора модели

# Байесовская регуляризация

•Параметры $\theta$ стали частью вероятностной модели

•Совместное распределение $p(X, T, \theta)$ может быть представлено в следующем виде $p(X, T, \theta) = p(X, T|\theta)p(\theta)$

•У нас появился множитель $p(\theta)$, с помощью которого можно ограничить множество допустимых значений $\theta$ и/или ввести предпочтения на те или иные значения

•Учитывая, что функция $p(X, T|\theta)$ нам была известна, задача выбора модели свелась к определению априорного распределения $p(\theta)$

•Введение ограничений на возможные значения настраиваемых в ходе обучения параметров, неиндуцированных обучающей выборкой часто называют регуляризацией

# Примеры задач выбора модели

•Определение числа кластеров в данных

•Выбор коэффициента регуляризации в задаче машинного обучения (например, коэффициента затухания весов (weight decay) в нейронных сетях)

•Установка степени полинома при интерполяции сплайнами

•Выбор наилучшей ядровой функции в методе опорных векторов (SVM)

•Определение количества ветвей в решающем дереве

•и многое другое...

Наивный байесовский классификатор (вики)

Пример и роль в современном мире