Лекция 6

1. Распознавание образов в условиях параметрической неопределенности на основе обучения с учителем

До сих пор мы рассматривали решение задач статистической теории распознавания образов при наличии полных статистических описаний классов образов, задаваемых, как известно, априорными вероятностями классов и их условными плотностями распределения (функциями правдоподобия) $\omega_i : p(\omega_i),\ p(x/\omega_i),\ i=\overline{1,M}$. На практике часто бывает, что данные характеристики классов отсутствуют и имеются лишь достаточно общие представления относительно вида и характера действующих статистических закономерностей.

Пусть в рамках постановки задачи машинного обучения с учителем для каждого класса имеется индивидуальная обучающая выборка данных $X^{N_i} = \{x^{(i,1)},...,x^{(i,N_i)}\}$, $x^{(i,s)} = (x_1^{(i,s)},...,x_n^{(i,s)})^T \in R^n$, $i=\overline{1,M}$. Обозначим также общую обучающую выборку $X^N = \bigcup_{i=1}^{M} X^{N_i}$.

Общим подходом для решения задачи в этой ситуации является проведение оценивания неизвестных параметров и характеристик с последующим использованием полученных результатов при синтезе алгоритмов распознавания.

Что касается оценки априорных вероятностей классов, то здесь эта задача может быть решена достаточно просто, если имеются основания предполагать, что относительные частоты появления объектов различных классов достаточно достоверно отражены в обучающих данных. Тогда можно считать, что $\tilde{p}(\omega_i) = N_i/N$, $N = N_1+,...,N_M$. К такому предположению, тем не менее, следует относиться осторожно и нужно его подкреплять определенными соображениями, вытекающими из предметного анализа ситуации. При отсутствии возможности обоснованного определения априорных вероятностей классов от их использования можно вообще отказаться и перейти к реализации стратегии максимального правдоподобия не требующей их использования.

Остается проблема задания функций правдоподобия классов. В этом случае используются методы и алгоритмы обработки информации, ориентированные на преодоление параметрической и непараметрической неопределенности статистических описаний объектов: а именно методы оценивания на основе обучающих данных неизвестных параметров плотности распределения

вероятностей при их известном аналитическом виде или методы оценивания плотности распределения в целом при отсутствии какой-либо информации об аналитической форме записи этой функции.

В рамках статистического подхода задача параметрического оценивания формулируется следующим образом. Известен общий вид условной плотности распределения вероятностей вектора наблюдений или признаков объектов данного класса $p(x/\omega_*)$, где ω_* – обозначает класс анализируемых объектов. При этом параметры $\theta=(\theta_1,...,\theta_h)^T$, определяющие точный вид этой плотности не известны. Обозначим этот факт следующим образом $p(x/\omega_*)=p(x/\omega_*,\theta)$. Далее использование обозначения класса, который является общим для всех наблюдений можно опустить $p(x/\omega_*,\theta)=p(x/\theta)$.

Требуется на основе полученной обучающей выборки $X^N=\{x^{(1)},...,x^{(N)}\}$, $x^{(i)}=(x_1^{(i)},...,x_n^{(i)})^T\in R^n$ определить неизвестные параметры в виде оценок $\tilde{\theta}=\tilde{\theta}(X^N)$, являющихся функциями полученных наблюдений. Это значение оценка является функций значений случайных величин и, следовательно, сама оценка является случайной величиной. Как случайная величина оценка $\tilde{\theta}=\tilde{\theta}(\mathbf{X}^N)$ обладает своей плотностью распределения $f_\theta(\theta)$.

В качестве примера такой ситуации можно привести распределение ГСВ с неизвестными математическим ожиданием и матрицей ковариации $p(x/\theta)=N(x,m,C)$, где вектор $\theta=(\theta_1,...,\theta_h)^T$ содержит в качестве компонентов n компонентов математического ожидания и $n\times(n-1)/2$ компонентов матрицы ковариации (с учетом ее симметричности). Тогда $h=n+n\times(n+1)/2$.

Вводятся следующие свойства оценок: несмещённость, состоятельность, эффективность и робастность, которые анализируются при построении алгоритмов оценивания. Выполнение указанных свойств характеризует получаемую оценка как «хорошую».

1.Оценка называется несмещенной, если выполняется $M[\tilde{\theta}(\mathbf{X}^N)]=M[\theta]$ (совпадают математические ожиданий оценки и оцениваемого вектора параметров). Оценка называется асимптотически несмещенной, если выполняется $\lim\limits_{N\to\infty}M[\tilde{\theta}(\mathbf{X}^N)]=M[\theta]$. Смещением оценки называется разность $\Delta\tilde{\theta}=M[\tilde{\theta}(\mathbf{X}^N)]-M[\theta]$.

2. Оценка называется состоятельной, если для любого $\varepsilon>0$ выполняется $\lim\limits_{N\to\infty}P[\|\tilde{\theta}(\mathbf{X}^N)-\theta\|]<\varepsilon$, т.е. имеет место сходимость по вероятности оценки и оцениваемой величины (вектора).

3. Оценка (несмещенная) называется эффективной, если для любой другой оценки $\hat{\boldsymbol{\theta}}(\mathbf{X}^N)$ выполняется $M[\|\tilde{\boldsymbol{\theta}}(\mathbf{X}^N)-\boldsymbol{\theta}\|^2] \le M[\|\hat{\boldsymbol{\theta}}(\mathbf{X}^N)-\boldsymbol{\theta}\|^2]$. Для одномерного случая данное свойство фактически означает, что дисперсия эффективной оценки меньше, чем у любой другой оценки.

4. Робастность оценки – свойство характеризующее устойчивость оценки по отношению к аномальным наблюдениям («выбросам»), возможно присутствующим в используемой для получения оценки выборке.

В теории оценивания существует много различных подходов. Следует выделить следующие подходы: подход, основанный на методе максимального правдоподобия; байесовский подход; метод моментов, метод интервального оценивания и др. Все они направлены на то, чтобы получить оптимальную в заданном смысле оценку неизвестных параметров и подробно рассмотрены в соответствующей литературе. Мы рассмотрим два первых подхода.

Далее в наших рассуждениях часто будет использоваться обозначение для значения векторного параметра $\tilde{\boldsymbol{\theta}}$, которое соответствует максимуму или минимуму функции $F(\boldsymbol{\theta})$ относительно этого параметра как ее аргумента или одного из аргументов. Мы будем использовать для этого следующие обозначения: для максимума $\tilde{\boldsymbol{\theta}} = \arg\max F(\theta)$ или $\tilde{\boldsymbol{\theta}} = \underset{\theta}{\operatorname{argmax}}\, F(\mathbf{X},\theta)$ или $F(\boldsymbol{\theta}) \to \max\limits_{\theta}$; для минимума $\tilde{\boldsymbol{\theta}} = \arg\min F(\theta)$ или $\tilde{\boldsymbol{\theta}} = \underset{\theta}{\operatorname{argmin}}\, F(\mathbf{X},\theta)$ или $F(\boldsymbol{\theta}) \to \min\limits_{\theta}$.

1. Метод максимального правдоподобия

При использовании метода максимального правдоподобия оцениваемые параметры рассматриваются как фиксированные и неизвестные. Наилучшая оценка определяется как величина, при которых вероятность реально наблюдаемых данных максимальна. При получении обучающей выборки наблюдений $X^N = \{x^{(1)},\ldots,x^{(N)}\}$ на основе плотности распределения известного вида $p(x/\boldsymbol{\theta})$ строят функцию правдоподобия выборки $P(X^N/\theta)$, которая с учетом независимости элементов обучающей выборки имеет вид

$$P(X^N/\boldsymbol{\theta}) = \prod_{i=1}^{N} p(x^{(i)}/\boldsymbol{\theta}).$$

Задача состоит в том, чтобы на основе полученных наблюдений максимизировать значения функции правдоподобия относительно значений вектора неизвестных постоянных параметров, т.е. найти оценку как

$$\tilde{\boldsymbol{\theta}} = \underset{\theta}{\arg\max}\, P(X^N/\boldsymbol{\theta}).$$

Для целей анализа обычно удобнее иметь дело с логарифмом функции правдоподобия, который имеет максимум в той же точке, что и исходная функция. В случае выполнения необходимых свойств дифференцируемости задача сводится к решению системы уравнений, в которой количество уравнений равно количеству определяемых параметров – компонентов вектора $\boldsymbol{\theta}$

$$\left.\frac{\partial \ln P(X^N/\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} = \left(\frac{\partial \ln P(X^N/\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1},\ldots,\frac{\partial \ln P(X^N/\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_h}\right)^T = \sum_{i=1}^{N}\frac{\partial \ln p(x^{(i)}/\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0},$$

$$\left.\frac{\partial \ln P(X^N/\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}\right|_{\boldsymbol{\theta}_j=\tilde{\boldsymbol{\theta}}_j} = \sum_{i=1}^{N}\frac{\partial \ln p(x^{(i)}/\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j} = \sum_{i=1}^{N}\frac{1}{p(x^{(i)}/\boldsymbol{\theta})}\frac{\partial p(x^{(i)}/\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j} = 0,\ \ j=\overline{1,h}$$

Следует отметить, что полученные таким образом решения определяют только необходимые условия экстремума функции правдоподобия и, часто, ими можно ограничиться. В тоже время, иногда, дополнительно проводится исследование условий достаточности при нахождении максимума.

Оценки постоянных параметров методом максимального правдоподобия (МП) обладают хорошими свойствами – они во многих случаях асимптотически состоятельны, эффективны и распределены по гауссовскому закону, что существенно упрощает их анализ.

Рассмотрим несколько примеров. Пусть неизвестным вектором параметров закона распределения $p(x/\boldsymbol{\theta}) = N(x,m,C)$ является вектор математического ожидания для ГСВ $\boldsymbol{\theta} = m$. Тогда

$$p(x^{(i)}/\boldsymbol{\theta}) = N(x^{(i)},m,C) = \frac{1}{(2\pi)^{n/2}|C|^{1/2}}\exp\left[-\frac{1}{2}(x^{(i)}-m)^T C^{-1}(x^{(i)}-m)\right],$$

$$\frac{\partial \ln p(x^{(i)}/\boldsymbol{\theta})}{\partial m_j} = -\frac{1}{2}\frac{\partial}{\partial m_j}\left[\sum_{k=1}^{n}\sum_{t=1}^{n}(x_k^{(i)}-m_k)\,\bar{c}_{kt}(x_t^{(i)}-m_t)\right] = -\sum_{k=1}^{n}\bar{c}_{jk}(x_k^{(i)}-m_k),\ \ j=\overline{1,n}.$$

Иначе, в векторном виде

$$\frac{\partial \ln p(x^{(i)}/\boldsymbol{\theta})}{\partial m} = -C^{-1}(x^{(i)}-m).$$

Отсюда получим окончательное векторное уравнение для нахождения оценки

$$\left.\frac{\partial \ln P(X^N/\theta)}{\partial m}\right|_{m=\tilde{m}} = -\sum_{i=1}^{N} C^{-1}(x^{(i)}-\tilde{m}) = 0$$

и, в итоге, оценку в виде

$$\tilde{m} = \frac{1}{N}\sum_{i=1}^{N} x^{(i)} .$$

Данная оценка в точности равна среднему арифметическому элементов выборки (выборочному среднему) и, как показано в работах по математической статистике, является несмещенной, состоятельной и эффективной. В частности, простейший анализ показывает, что матрица ковариации ошибки оценивания в данном случае равна

$$R_\varepsilon = M[(\tilde{\mathbf{m}}-\mathbf{m})(\tilde{\mathbf{m}}-\mathbf{m})^T] = \frac{1}{N}C .$$

Более сложным является случай, когда неизвестными параметрами являются как вектор математического ожиданий, так и матрица ковариации. Решение базируется на проведении, по сути, аналогичных, но более громоздких выкладок. Оценки по максимуму правдоподобия в этом случае даются выражениями

$$\tilde{m} = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}, \quad \tilde{C} = \frac{1}{N}\sum_{i=1}^{N}(x^{(i)}-\tilde{m})(x^{(i)}-\tilde{m})^T,$$

т.е. являются выборочным средним и выборочной матрицей ковариации. Так как подлинная матрица ковариации есть математическое ожидание $C = M[(\mathbf{x}-m)(\mathbf{x}-m)^T]$, то полученный результат вполне очевиден. Следует также отметить, что полученная методом МП оценка матрицы ковариации являются смещенной. Несмещенная оценка задается соотношением

$$\hat{C} = \frac{1}{N-1}\sum_{i=1}^{N}(x^{(i)}-\tilde{m})(x^{(i)}-\tilde{m})^T .$$

Очевидно, что разница между этими оценками стирается при больших объемах обучающей выборки.

Пусть неизвестным параметром биномиального закона распределения при проведении N испытаний является вероятность появления события $\theta = p$. Здесь элементами выборки являются значения случайной величины $\mathbf{x}$: $X^N = \{x^{(1)},\ldots,x^{(N)}\}$, каждое из которых может принимать два значения: 1 – при появлении события в данном испытании и 0 – в противном случае. Тогда

$$p(X^N/\theta) = C_N^u p^u (1-p)^{N-u},\ u = \sum_{i=1}^{N} x^{(i)},$$

$$\frac{d \ln p(X^N/\theta)}{dp} = C_N^u \left(\frac{u}{p} - \frac{N-u}{1-p} \right).$$

Отсюда получим уравнение для нахождения оценки

$$\left. \frac{d \ln P(X_N/\theta)}{dp} \right|_{p=\tilde{p}} = C_N^u \left(\frac{u - u\tilde{p} - \tilde{p}N + \tilde{p}u}{\tilde{p}(1-\tilde{p})} \right) = 0$$

и, в итоге, оценку в виде

$$\tilde{p} = \frac{u}{N}.$$

С учетом того, что $m_u = p$, $D_u = p(1-p)/N$ данная оценка является несмещенной, состоятельной и эффективной. Таким образом, получен очевидный результат: оценкой максимального правдоподобия вероятности события является отношение частоты его появления в серии из N испытаний к числу этих испытаний.

2. Байесовское оценивание

В основе байесовского подхода к оцениванию параметров лежит использование представления о неизвестном параметре как о значении случайной величины θ с заданным априорным законом распределения $f_\theta(\theta) = p(\theta)$. Кроме того, вводится функция риска (функция потерь), определяющая стоимость ошибок оценивания в каждом конкретном опыте, – $r[\theta, \hat{\theta}(X^N)]$, где $\hat{\theta}(X^N)$ некоторая, не обязательно оптимальная, оценка. Вычисляется средний риск

$$R(\hat{\theta}) = \int r[\theta, \hat{\theta}(X^N)] P(\theta, X^N) d\theta dX^{(N)} = \int_{X^N} \{ \int_{\theta} r[\theta, \hat{\theta}(X^N)] p(\theta/X^N) d\theta \} P(X^N) dX^N .$$

В данной постановке задачи считается, что оптимальная оценка соответствует минимуму среднего риска. Внутренний интеграл в называется апостериорным риском. Средний риск будет минимальным, если для каждой конкретной реализации X^N будет минимизироваться апостериорный риск. Это можно сделать, так как оценка $\hat{\theta}(X^N)$ определяется для каждой совокупности наблюдений X^N.

В результате получается, что минимум среднего риска достигается при том же значении $\hat{\boldsymbol{\theta}}(X^N)$, что и минимум апостериорного риска, которое и будет оптимальной оценкой $\tilde{\boldsymbol{\theta}}(X^N)$,

$$R(\theta / X^N) = \int_{\theta} r[\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}(X^N)] p(\theta / X^N) d\theta \to \min_{\tilde{\theta}}.$$

Для получения конкретной оценки следует далее использовать конкретную функцию потерь. Чаще всего используют три типичные функции потерь: квадратичную, простую и модульную

$$r[\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(X^N)] = \left\| \theta - \hat{\boldsymbol{\theta}} \right\|^2 = (\theta - \hat{\boldsymbol{\theta}})^T S (\theta - \hat{\boldsymbol{\theta}}), \quad r[\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(X^N)] = 1 - \boldsymbol{\delta}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}), \quad r[\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(X^N)] = \left\| \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \right\|.$$

где $S \geq \mathbf{0}$ – симметричная неотрицательно определенная матрица, например единичная $S = I$.

При использовании простой функции потерь минимум апостериорного риска достигается, когда оценка соответствует максимуму апостериорной плотности распределения вероятностей (иногда говорят максимума апостериорной вероятности (МАВ)) оцениваемого параметра

$$R(\boldsymbol{\theta} / X^N) = 1 - \int_{\theta} \boldsymbol{\delta}[\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}(X^N)] p(\boldsymbol{\theta} / X^N) d\boldsymbol{\theta} = 1 - p(\tilde{\boldsymbol{\theta}} / X^N) \to \min_{\tilde{\theta}},$$

$$\tilde{\boldsymbol{\theta}}(X^N) : p(\boldsymbol{\theta} / X^N) = \frac{P(X^N / \boldsymbol{\theta}) p(\boldsymbol{\theta})}{P(X^N)} \to \max_{\theta}.$$

В качестве примера решения задачи оценивания в рамках байесовского подхода рассмотрим ту же самую задачу определения математического ожидания многомерного гауссовского распределения $p(x / \boldsymbol{\theta}) = N(x, m, C)$ с известной матрицей ковариации на основе метода максимума апостериорной плотности вероятностей, реализуемого при простой функции потерь. Пусть априорное распределение оцениваемого вектора математического ожидания для ГСВ ($\boldsymbol{\theta} = \mathbf{m}$) также является гауссовским

$$p(m) = N(m, m_0, C_0).$$

Тогда апостериорная плотность распределения можно представить в виде

$$p(\theta = m / X^N) = \frac{P(X^N / \theta) p(\theta)}{P(X^N)} = \alpha_1 \prod_{i=1}^{N} p(x^{(i)} / m) p(m) = \alpha_1 \prod_{i=1}^{N} N(x^{(i)}, m, C) N(m, m_0, C_0),$$

где $\alpha_1 = [P(X^N)]^{-1}$ – нормировочный множитель, зависящий от X^N. Перепишем последнее выражение, выделим в экспоненте слагаемые при m следующим образом:

$$p(m / X^N) = \alpha_2 \exp\left\{-\frac{1}{2}\left[m^T (NC^{-1} + C_0^{-1}) m - 2 m^T (C^{-1} \sum_{i=1}^{N} x^{(i)} + C_0^{-1} m_0)\right]\right\},$$

где α_2 – нормировочный множитель, не зависящий от m. Апостериорную плотность можно представить в виде

$$p(m / X^N) = \alpha_3 \exp[-\frac{1}{2}(m - m_N)^T C_N^{-1} (m - m_N)],$$

где $\alpha_3 = [(2\pi)^{n/2} |C_N|^{1/2}]^{-1}$ – очередной нормировочный множитель для гауссовской плотности распределения.

Приравнивая матричные коэффициенты

$$NC^{-1} + C_0^{-1} = C_N^{-1}, \quad NC^{-1}\tilde{m} + C_0^{-1} m_0 = C_N^{-1} m_N,$$

где $\tilde{m} = N^{-1} \sum_{i=1}^{N} x^{(i)}$ – уже известное нам выборочное среднее, получим

$$C_N = (NC^{-1} + C_0^{-1})^{-1} = C_0 (C_0 + \frac{1}{N} C)^{-1} \frac{1}{N} C,$$

$$m_N = C_N (NC^{-1}\tilde{m} + C_0^{-1} m_0) = C_0 (C_0 + \frac{1}{N} C)^{-1} \tilde{m} + \frac{1}{N} C (C_0 + \frac{1}{N} C)^{-1} m_0.$$

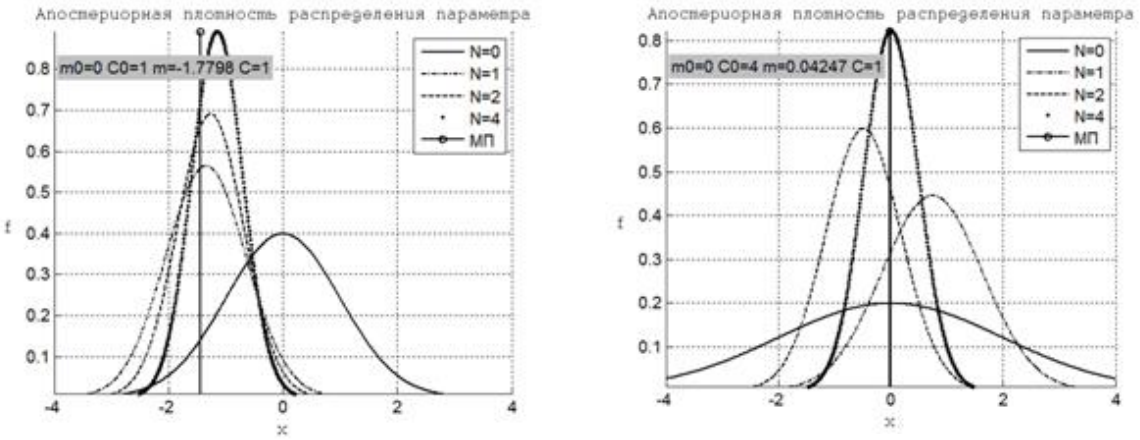
При получении окончательных выражений использовались известные матричное тождество, справедливое для двух невырожденных матриц A, B

$$(A^{-1} + B^{-1})^{-1} = A(A + B)^{-1} B = B(A + B)^{-1} A.$$

Анализ полученных выражений показывает, что при увеличении N получаемая плотность распределения становится дельтаобразной (сужается), при этом роль параметров априорного распределения становится менее существенной.

В качестве оценки МАВ здесь выступает величина m_N, а в качестве матрицы ковариации ошибки $R_\varepsilon = M[(\tilde{m} - m)(\tilde{m} - m)^T]$ матрица C_N.

Для проведения анализа полученного алгоритма по методу МАВ и его сравнения с алгоритмом на основе МП рассмотрим иллюстративный пример с использованием разработанной программы для исследования алгоритмов оценки математического ожидания гауссовской случайной величины. В этой программе первоначально строятся графики (рис. 1) для априорной и апостериорной плотностей распределения оцениваемого параметра для различных значений N, определяющих объемы обучающей выборки.



а) б)

Рис. 1. Типовые графики апостериорных плотностей распределения параметра для различных априорных плотностей

Оценка МП на этих графиках отображается в виде вертикальной прямой линии и соответствует случаю максимального значения $N = 4$. Оценка МАВ соответствует максимуму финальной апостериорной плотности распределения для того же N. Одновременно указанные графики иллюстрируют процесс эволюции апостериорной плотности распределения оцениваемого параметра при увеличении объема анализируемых данных.

Во второй части программы проводится сравнение оценок МП и МАВ по величине дисперсии ошибки оценивания для различных соотношений априорной дисперсии C_0 и дисперсии плотности распределения наблюдаемой случайной величины C при фиксированном значении $N = 8$. Аналитические соотношения для дисперсии получены на основе ранее приведенных формул для общего случая оценки случайного вектора. Экспериментальные оценки получены путем статистического моделирования и усреднения 10000 реализаций процедуры проведения оценивания. Результаты представлены в табл. 1.

Таблица 1. Результаты определения дисперсии ошибок оценивания

Априорная и	Оценка МП		Оценка МАВ	

апостериорная дисперсии распределения параметра	Аналитический расчет	Статистическое моделирование	Аналитический расчет	Статистическое моделирование
$C_0 = \mathbf{1}, C = \mathbf{1}$	0.1250	0.1258	0.1111	0.1121
$C_0 = \mathbf{4}, C = \mathbf{1}$	0.1250	0.1260	0.1212	0.1215
$C_0 = \mathbf{0.25}, C = \mathbf{1}$	0.1250	0.1237	0.0833	0.0827
$C_0 = \mathbf{0.1}, C = \mathbf{1}$	0.1250	0.1248	0.0556	0.0562

Их анализ показывает, что при сужении априорной плотности распределения оценки МАВ позволяют получить ощутимый выигрыш по сранению с оценками МП и, наоборот, при расширении области локализации априорного распределения параметра этот выигрыш нивелируется.

3. Использование оценок максимального правдоподобия и байесовских оценок в подстановочных алгоритмах

При проведении синтеза алгоритма распознавания на основе МП используются так называемые подстановочные алгоритмы (plug-in). Идея этих алгоритмов достаточно проста: полученные оценки подставляются в выражения для плотностей распределения вместо неизвестных параметров $\tilde{p}(x/\omega_i) = p(x/\tilde{\theta}_i, \omega_i)$, $i = \overline{1,M}$, которые затем и используются при получении алгоритмов принятия решения известного вида или других его разновидностей.

В случае двух классов общая структура решающего правила имеет вид

$$\frac{N_1}{N_1+N_2}\tilde{p}(x/\omega_1) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{N_1}{N_1+N_2}\tilde{p}(x/\omega_2),$$

где $\tilde{p}(x/\omega_i) = p(x/\tilde{\theta}_i, \omega_i)$ – плотность, полученная путем подстановки оценки параметра, в случае использования МП, или $\tilde{p}(x/\omega_i) = p(x/X^{N_i}, \omega_i)$ – плотность, полученная путем выполнения интегрального преобразования ,в случае использования МАВ.

Полученные оценки могут также подставляться в выражения вида для получения условных относительно реализаций $X^{N_i}, x^{(i,s)} = (x_1^{(i,s)}, \ldots, x_n^{(i,s)})^T$, $i = \overline{1,M}$ вероятностей ошибок распознавания. Безусловные вероятности ошибок могут быть получены путем усреднения получаемых вероятностей относительно реализаций обучающих данных.

В асимптотике при $N_i \to \infty$ получаемые оценки обладают рядом оптимальных свойств: они не смещены, состоятельны и эффективны. Это дает основание надеяться, что получаемые после подстановки оценок неизвестных параметров решающие правила будут также близки по своим вероятностным характеристикам к синтезированным при полной априорной неопределенности. Однако в условиях конечных, зачастую малых, выборок асимптотические свойства не гарантируют точного восстановления функций правдоподобия, и, следовательно, высокого качества распознавания. Кроме того, использование правил подстановки лучше всего работает для гауссовских распределений или распределений схожего вида. В случае если классы распадаются на изолированные сгустки или в компонентах вектора признаков имеются дискретные признаки качество классификации таких алгоритмов резко падает.

Из всех этих соображений следует, что принципиальным вопросом для подстановочных алгоритмов является анализ вероятностей ошибок в зависимости от объема обучающих данных. При этом могут использоваться как ранее полученные аналитические соотношения с использованием полученных оценок параметров, которые в этом случае уже будут давать случайные результаты как функции выборки. Кроме того, всегда остается возможность получить данные прямого имитационного моделирования алгоритмов в виде относительных частот ошибок, которые дают результаты, отражающие реально возникающие ситуации.

При использовании байесовских алгоритмов оценивания с целью обучения алгоритмов распознавания в условиях параметрической неопределенности в общем случае нет необходимости получения окончательных оценок. Достаточно определить апостериорную плотность и вычислить искомую функцию правдоподобия класса на основе следующего интегрального преобразования:

$$\tilde{p}(x/\omega_i) = \int p(x/\theta_i,\omega_i)p(\theta_i/X^{N_i})d\theta_i = \int p(x/\theta_i,\omega_i)\frac{p(X^{N_i}/\theta_i)p(\theta_i)}{P(X^{N_i})}d\theta_i .$$

Тем самым, фактически получены точные выражения для условных относительно наблюдаемых данных плотностей распределения классов $\tilde{p}(x/\omega_i) = p(x/X^{N_i},\omega_i)$.

Рассмотрим, например, задачу обучения классификатора с гауссовскими функциями правдоподобия при неизвестных математических ожиданиях. Если использовать полученные выше результаты для апостериорной

плотности распределения неизвестного математического ожидания гауссовского многомерного распределения

$$p(\theta_i = m_i / X^{N_i}) = N(m_i, m_{N_i}, C_{N_i}),$$

и воспользоваться формулой интегрального преобразования условных гауссовских плотностей получим

$$\tilde{p}(x/\omega_i) = \int N(x, m_i, C_i) N(m, m_{N_i} C_{N_i}) dm = N(x, m_{N_i}, C_i + C_{N_i}).$$

Т.е. в данном случае параметризация плотностей распределения классов получается абсолютно точно и полученные выражения можно использовать при получении алгоритмов принятия решения. При этом значения $m_{N_i}, \tilde{C}_i = C_i + C_{N_i}$ можно подставить в выражения для получения условных относительно реализаций X^{N_i}, $i = \overline{1,M}$ вероятностей ошибок распознавания. Безусловные вероятности ошибок также могут быть получены путем усреднения получаемых вероятностей относительно реализаций обучающих данных.

Для большинства плотностей $p(x/\omega_i, \theta_i)$ вид апостериорной плотности $p(\theta_i / X^{N_i}, \omega_i)$ сходится к дельта-функции при $N_i \to \infty$ относительно оценки МАВ – $\tilde{\theta}_i$. Это, в частности, означает, что

$$\tilde{p}(x/\omega_i) \approx \int p(x/\theta_i, \omega_i) \delta(\theta_i - \tilde{\theta}_i) d\theta_i = p(x/\tilde{\theta}_i, \omega_i).$$

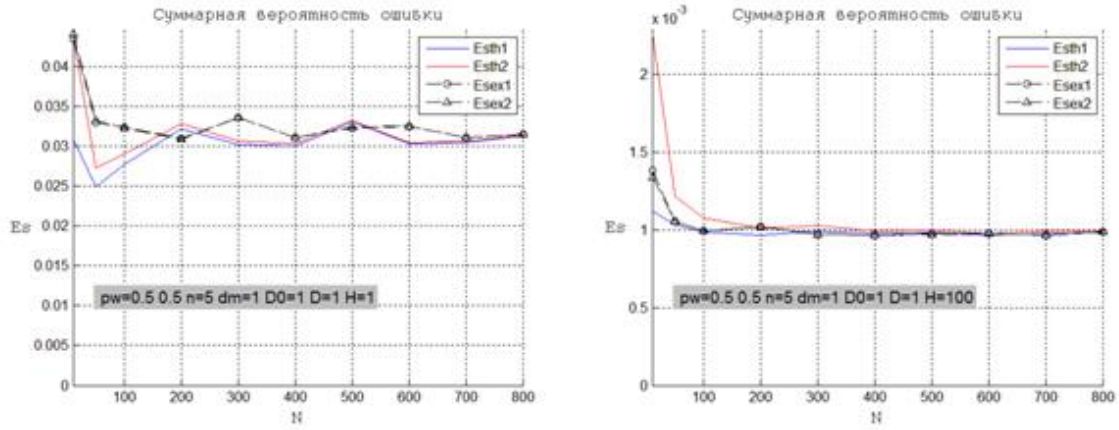
Т.е. в данном случае также фактически реализуется подстановочный алгоритм, причем, если еще и априорная плотность параметра является достаточно ровной в окрестности получаемой оценки, то результаты обучения с использованием методов МП и МАВ совпадают.

Главная проблема применения байесовского подхода состоит в сложности выполнения представленных выше интегральных преобразований и получения апостериорных плотностей распределения неизвестных параметров. В общем многомерном случае возникают существенные трудности, которые ограничивают использования статистически оптимальных оценок параметров как общего подхода к решению задачи обучения с учителем.

При проведении анализа алгоритмов, прежде всего, вызывает интерес получение зависимостей качества распознавания от объема обучающих

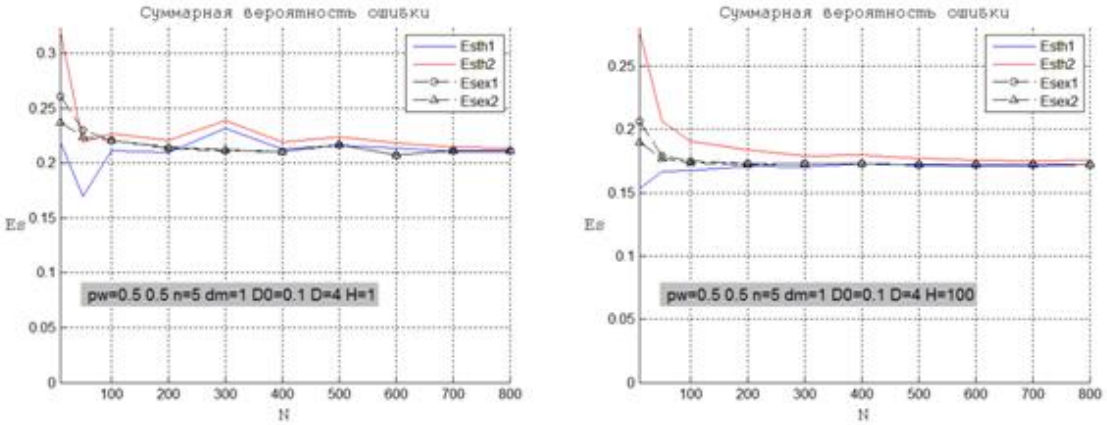
данных и сравнение результатов, получаемых на основе МП и МАВ.

На рис. 2 и 3 представлена результаты работы программы для случая $n = 5$ при различных значениях количества испытаний H, направленных на усреднение суммарной вероятности ошибки, а также при различных соотношений D_0 (D0) и D(D), определяющих дисперсию априорного распределения компонент математических ожиданий и дисперсию распределения компонент вектора признаков. Зависимости обозначены следующим образом: $Esth1$ и $Esth2$ –усредненные оценки суммарной вероятности ошибки, полученные с использованием аналитических соотношений, и затем усредненные по числу испытаний процесса обучения; $Esex1$ и $Esex2$ – дают оценки суммарной вероятности ошибки, полученные исключительно в ходе статистического моделирования. На рис.2а и 2а значение равно $H = 1$, а на рис.2б и 3б – $H = 100$.



а) б)

Рис.2. Зависимости для суммарной вероятности ошибки от объема обучающих выборок при D0=1 и D=1



а) б)

Рис.3. Зависимости для суммарной вероятности ошибки от объема обучающих выборок при D0=0.1 и D=4

Анализ представленных зависимостей показывает, что вероятности ошибок $\tilde{\alpha}(X^N), \tilde{\beta}(X^N)$ и, в итоге, суммарная вероятность ошибки всех алгоритмов при однократном обучении являются случайными величинами, которые зависят от использованных реализаций обучающих данных. После усреднения зависимости приобретают достаточно гладкий характер, что позволяет увидеть близость полученных оценок к общему предельному значению уже при объеме обучающих выборок, превышающем $N_{1,2} > 400$. Разница между алгоритмами МП и МАВ с точки зрения их влияния на результат распознавания в рассмотренных примерах незначительна и заметна только при малых D_0 и $N_{1,2}$.

◄ Лекция 5
Перейти на...
Перейти на...
Лекция 7 ►