Линейная классификация. Градиентные методы.

Московский физико-технический институт, МФТИ

Москва

Основные понятия и обозначения

Дано: выборка обучающих пар обьектов $X^l = (x_i, y_i)_{i=1}^l$.
В общем виде алгоритм классификации представим функцией $a(x, w) = \operatorname{sign} f(x, w)$. *Задача*: найти разделяющую поверхность $f(x, w) = 0$.
Отсутпом обьекта называется величина $M_i(w) = y_i f(x_i, w)$ относительно алгоритма классификации $a(x, w)$.

Аппроксимация эмпирического риска

Пусть - монотонно невозрастающая функция отсупа, мажорирующую функцию потерь $[M < 0] \leq \mathcal{L}(M)$.

$$Q(w, X^l) = \sum_{i=1}^{l} [M_i(w) < 0] \leq \tilde{Q}(w, X^l) = \sum_{i=1}^{l} (M_i(w)) \to \min_{w} .$$

$Q(M) = (1 - M)^2$ - квадратичная,

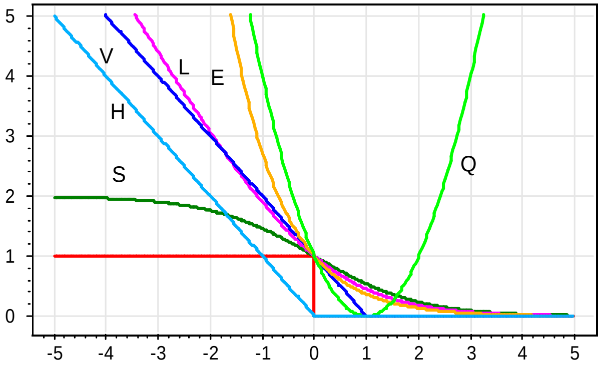
$V(M) = \max\{0, 1 - M\}$ - кусочно-линейная,

$S(M) = 2/(1 + e^{-M})$ - сигмоидная,

$L(M) = \log_2(1 + e^{-M})$ - логистическая,

$E(M) = e^{-M}$ - экспоненциальная.

Аппроксимация пороговой функции



Идея персептрона

- Мак-Каллока в 1943 впервые представили идею использования нейронных сетей в качестве вычислительных машин;
- Хебба в 1949 впервые ввел правило самоорганизующегося обучения;
- Розенблатт в 1958 году ввел понятие переспетрона как первой модели обучения с учителем.

Мат модель линейной классификации

Рассмотрим классифицирующие модели вида $a(x, w) = \operatorname{sign} f(x, w)$, так что множество значений функционала $Y = \{-1, +1\}$. Функция доли неправильных ответов

$$Q(a, x) = \frac{1}{l} \sum_{i=1}^{l} [a(x_i) \neq y_i] = \frac{1}{l} \sum_{i=1}^{l} [\operatorname{sign}\langle w, x_i \rangle \neq y_i] \to \min_{w},$$

или в более компактной записи

$$Q(a, x) = \frac{1}{l} \sum_{i=1}^{l} [y_i \langle w, x_i \rangle <]$$

положим $M_i = y_i \langle w, x_i \rangle$.

Основные обозначения и определения

Введем обозначения:

$$x^* = \arg\min_{x \in X} f(x),\ X \subset H,\ (*)$$

$$f^* = f(x^*) = \min_{x \in X} f(x).$$

Теорема

Пусть X - компакт в H, тогда $f(x)$ - непрерывный X на функционал. Тогда существует точка глобального минимума $f(x)$ на X.

Основной итерационный процесс

Определим последовательность:

$$x^{n+1} = x^n + \alpha_n h^n,\ n = 0, 1, 2..$$

Обозначим основные этапы алгоритма отпмизации:

1. Положить $n = 0$, задать x^0;
2. Проверить условия останова;
3. Вычислить α_n;
4. Вычислить x^{n+1};
5. Увеличить на единицу. Перейти к п. 2;

Методы нулевого/первого и более порядков.

Критерии остановки

Могут применяться следующие критерии остановки процесса минимизации:

1. $$||x^{n+1} - x^*|| \leq \varepsilon_1,$$
2. $$|f(x^{n+1}) - f(x^n)| \leq \varepsilon_2,$$
3. $$|f'(x^n)| \leq \varepsilon_3.$$

Методы спуска

Пусть известно направление спуска такое что $f(x + \alpha x) < f(x)$. Пусть заданы x^n, h^n, необходимо выбрать α_n, такое что

$$f(x^n + \alpha_n h^n) = \min_{\alpha \geq 0} f(x^n + \alpha h^n).$$

Данную задачу не сложно решить в явном виде для квадратичного функционала:

$$f(x) = \frac{1}{2}(Ax, x) + (b, x) + (c, c).$$

Метод доверительной области

Рассмотрим приближенную модель с учетом ограниченного шага

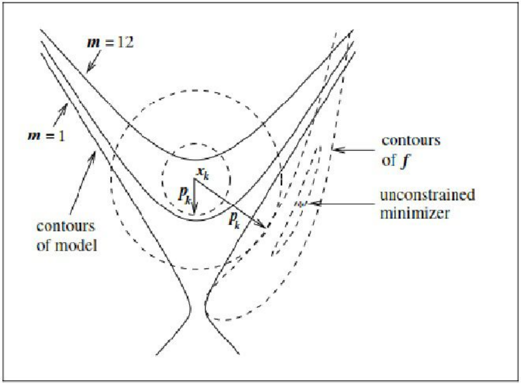
$$\min_{p \in \mathbb{R}} m_k(p) = \frac{1}{2} p^T B_k p + g_k^T p + f_k, \ ||p|| \leq \triangle_k,$$

важна величина близости модели к исходной функции

$$\rho_k = \frac{f(x_k) - f(x_k + p_k)}{m_k(0) - m_k(p_k)},$$

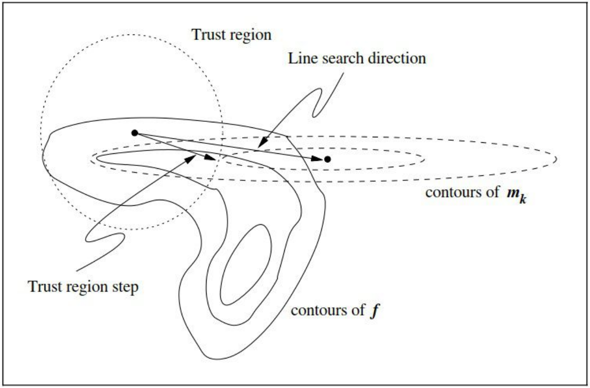
идеальный случай $\rho_k \sim 1$, если ρ_k - маленькое - уменьшаем область, если ρ_k близко к 1 и шаг p_k достигает границы - увеличиваем.

Контуры минимизации



Разница подходов:
доверительная область - максимальный радиус - направление
Для методов спуска - направление - длина шага

Контуры доверительной области



Методы спуска с условиями Вульфа и Голдштайна

Условие Вульфа

$$f(x_k + \alpha_k p_k) \leq f(x_k) + c_1 \alpha_k \nabla f_k^T p_k = l(\alpha),\ c_1 \in (0, 1),$$

в том числе условия кривизны

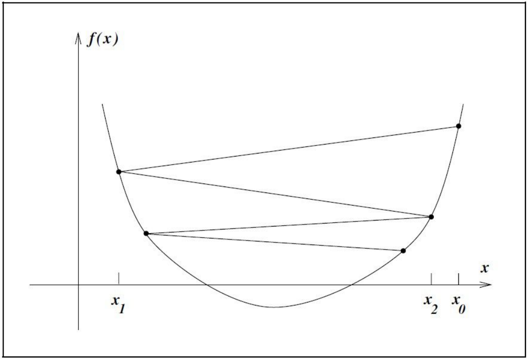
$$\nabla f(x_k + \alpha_k p_k)^T p_k \geq c_2 \nabla f_k^T p_k,\ c_2 \in (c_1, 1),$$

Условие Гольдштайна

$$f(x_k) + (1 - c)\alpha_k \nabla f_k^T p_k \leq f(x_k + \alpha_k p_k) \leq f(x_k) + c\alpha_k \nabla f_k^T p_k,$$

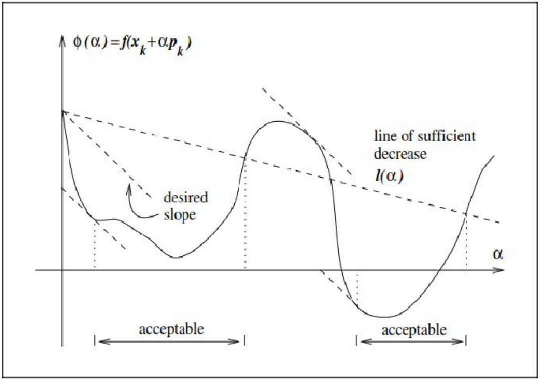
Контрпример

Рассмотрим последовательность $f(x_k) = 1/k$

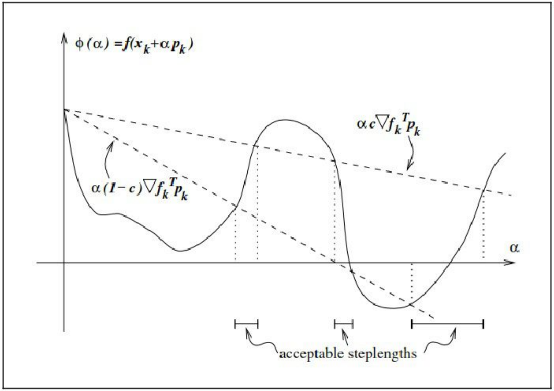


которая не сходится к минимуму $f(x^*) = -1$.

Условия Вульфа иллюстрация



Условия Гольдштайна, иллюстрация



Метод наискорейшего спуска

Пусть функционал имеет квадртоичный вид

$$f(x) = \frac{1}{2}x^T Qx - b^T x,$$

Матрица Q - симметрична и положительно определена, минимум соответствует решению уравнения $Qx = b$. Приравнивая к нулю производные функции $f(x_k - \alpha \nabla f_k)$, находим значение оптимального параметра

$$\alpha_k = \frac{\nabla f_k^T \nabla f_k}{\nabla f_k^T Q \nabla f_k},$$

получим итерационный процесс

$$x_{k+1} = x_k - \left(\frac{\nabla f_k^T \nabla f_k}{\nabla f_k^T Q \nabla f_k} \right) \nabla f_k.$$

Алгоритм Бройдена — Флетчера — Гольдфарба — Шанно (BFGS)

Пусть получена дискретизация функционала

$$m_k(x) = \frac{1}{2} p^T B_k p + \nabla f_k^T p + f_k,$$

где вектор p используется в качестве направления спуска $x_{k+1} = x_k + \alpha_k p_k$ (α_k- можно найти например по методу Вульфа). Построим итерационный процесс для обновления матрицы Гессе:

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}.$$

Левенберга - Марквардта метод

Связь метода с методом доверительной области

$$\min_p \frac{1}{2}||J_k p + r_k||^2, \; ||p|| \leq \triangle_k,$$

когда достиагется граничное условие, задачу можно свести к

$$(J^T J + \lambda I)p = -J^T r.$$

Общий вид градиентного метода

Функци эмпирического риска

$$Q(w) = \sum_{i=1}^{l} \mathcal{L}_i(w) \to \min_{w}.$$

Итерационный градиентный метод в общем виде

$$w^{(t+1)} = w^{(t)} - h \sum_{i=1}^{l} \mathcal{L}_i(w^{(t)}).$$

Пусть задано $w^{(0)}$ - начальное приближение.
и h - темп обучения. Для линейной модели

$$w^{(t+1)} = w^{(t)} - h \sum_{i=1}^{l} \mathcal{L}_i'(\langle w^{(t)}, x_i \rangle y_i) x_i y_i.$$

Метод стохаостического градиентов

Входные параметры: темп обучения h, темп забывания λ.

1. инициализация веса w_j, $j = 0..n$;
2. инициализация невязки $\bar{Q} = \frac{1}{l}\sum_{i=1}^{l} \mathcal{L}_i(w)$;
3. выбираем обьект x_i из X^l случайным образом;
4. вычисляем потерю $\varepsilon_i = \mathcal{L}_i(w)$;
5. итерация по шагам $w_{n+1} = w_n - h\nabla\mathcal{L}_i(w)$;
6. оценка функционала $Q = \lambda\varepsilon_i + (1-\lambda)\bar{Q}$
7. критерий остановки: значения $\bar{Q}$ и веса w не сойдутся.

Диагональный метод Левенберга-Марквардта

Методы типа Ньютоновских (Ньютона-Рафсона)

$$w_{n+1} = w_n = h(\mathcal{L}_i''(w))^{-1}\nabla\mathcal{L}_i(w),$$

где $\mathcal{L}_i''(w) = \left(\frac{\partial^2\mathcal{L}_i(w)}{\partial w_j \partial w_{j'}}\right)$. По аналогии с Левенбергом-Марквардтом:

$$w_{n+1} = w_n = h\left(\frac{\partial^2\mathcal{L}_i(w)}{\partial w_j^2} + \mu\right)^{-1}\nabla\frac{\mathcal{L}_i(w)}{\partial w_j},$$

отношение h/μ - темп обучения на ровных участках функционала $\mathcal{L}_i(w)$, где вторая производная обнуляется.

Вероятностная модель данных

Применим метод *максимума правдоподобия*. Пусть все наблюдения **независимы** , каждое из которых описывается функцией распределения $p(x, y|w)$, тогда правдопбие выборки можно представить $p(X^l|w) = \prod_{i=1}^{l} p(x_i, y_i|w) \to \max_w$, так что указанный метод эквивалентен постановке минимизации ошибок или функции потерь

$$-\sum_{i=1}^{l} \ln p(x_i, y_i|w) = \mathcal{L}(y_i f(x_i, w)).$$

Регулярзиация

Введем априорное распределение параметров модели $p(w)$, так что по формуле условной вероятности плотность вероятности примет вид $p(x, y; \gamma) = p(x, y|w)p(w; \gamma)$. При этом в принципе максимума правдоподобия появится регуляризирующее слагаемое:

$$L_\gamma(w, X^l) = \ln p(X^l, w; \gamma) = \sum_{i=1}^{l} p(x_i, y_i|w) + lnp(w; \gamma) \to \max_w$$

Априорное распределение Гаусса и Лапласа

Распределение Гаусса соответствует квадратичной (L2) регуляризации

$$p(w; C) = \frac{1}{(2\pi C)^{n/2}} \exp\left(-\frac{||w||^2}{2C}\right).$$

Распределение Лапласа соответствует регуляризации первого порядка (L1)

$$p(w; C) = \frac{1}{(2C)^{n}} \exp\left(-\frac{||w||}{C}\right).$$

Где дисперсия $Dw_j = C$, а C - коэффициент регуляризации.

Базовые предоположения

Предположение

Множество прецедентов $X \times Y$ *является вероятностным пространством. Выборка прецедентов* $X^l = (x_i, y_i)_{i=1}^l$ *получена случайно и независимо согласно вероятностному распределению с плотностью* $p(x, y) = P_y p_y(x) = P(y|x)p(x)$*, где* P_y*- априорные вероятности,* p_y *- функции правдоподобия,* $P(y|x)$ *- апостериорные вероятности классов* $y \in Y$*.*

Предположение

Функции правдопдобия классов принадлежат экспоненциальному семейству плотностей, имеют равные значения параметров d *и* δ*, но отличаются значениями параметров сдвига* θ_y*.*

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN},$$

$$precision = \frac{TP}{TP + FP},$$

$$recall = \frac{TP}{TP + FN},$$

Критерий качества на основе точности и полноты: F - мера, гармоническое среднее точности и полноты:

$$F = \frac{2 * precision * recall}{precision + recall}$$

Площадь под ROC -кривой

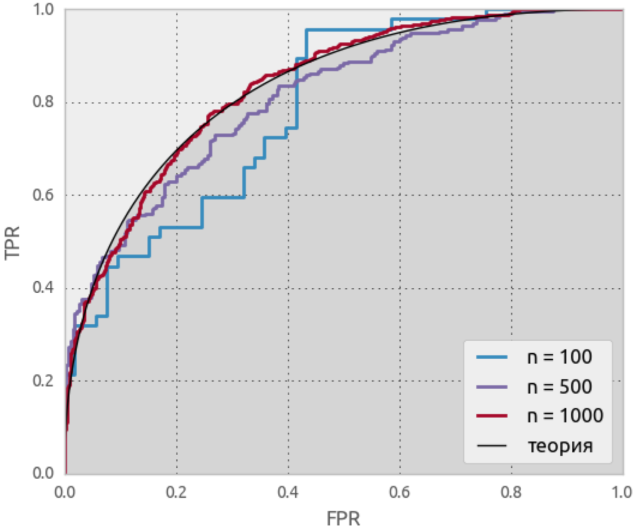
Доли неверно принятых обьектов (False Positive Rate) и верно принятых обьектов (True Positive Rate)

$$FPR = \frac{FP}{FP + TN},$$

$$FPR = \frac{TP}{TP + FN}.$$

Проблема чувствительности к соотношению классов -> решение через precision-recall кривая оценки.

Пример ROC -кривой



Дальнейшие темы связанные с классификацией

- Различные оценки точности алгоритма ROC-AUC кривые;
- Метод опорных векторов;
- Баесовские методы классификации;
- Многоклассовая классификация;