Лекция 5. Регуляризация, разреженные модели, линейная классификация.

Регуляризация. Замечания.

⋆ Нельзя включать w_0 в регуляризацию, так как он помогает выровнять масштабы признаков и целевой переменной, а также он никак не взаимодействует с признаками, поэтому нестрашно, если он неожиданно окажется большим.

⋆ Пусть есть признаки разных масштабов. Тогда будем штрафовать сильнее те веса, которые стоят при признаках другого масштаба. $||w||$ ведет себя неадекватно при немасштабированных признаках. То есть нужно масштабировать признаки перед регулризацией. (ускоряет градиентный спуск).

⋆ Добавление регуляризатора ускоряет GD, так как упрощает рельеф $Q(w)$. (график становится больше похож на параболу.)

Разреженные модели.

Зачем занулять часть весов?

1. Мы накидали в признаки все подряд.
2. Ускорение моделей.
3. $\ell \ll d$

Решение для линейных моделей: L_1-регуляризация. $Q(w) + \lambda||w||_1 \to \min_w$

Объяснение 1.

$$Q(w) + \lambda||w||_1 \to \min_w \quad \Leftrightarrow \quad \begin{cases} Q(w) \to \min_w \\ ||w||_1 \le C \end{cases}, \quad \text{для некоторого } C.$$

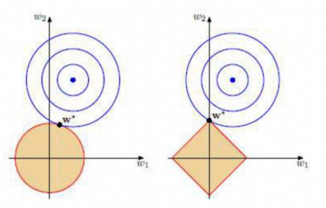

Рис. 1. Линии уровня функционала качества, а также ограничения, задаваемые L_2 и L_1-регуляризаторами.

Объяснение 2.

$w = \Big(1, \underbrace{\varepsilon}_{\sim 0{,}1}\Big)$, $0 < \delta < \varepsilon \ll 1$. Можно уменьшить один вес на δ, вопрос: какой выгоднее?

$$||w - (\delta, 0)||_2^2 = 1 - 2\delta + \delta^2 + \varepsilon^2, \qquad ||w - (\delta, 0)||_1 = 1 - \delta + \varepsilon$$

$$||w - (0, \delta)||_2^2 = 1 - 2\varepsilon\delta + \delta^2 + \varepsilon^2, \qquad ||w - (0, \delta)||_1 = 1 - \delta + \varepsilon$$

То есть, с точки зрения L_1 регуляризации неважно, какой коэффициент приближать к нулю.

А с точки зрения L_2 регудяризации выгоднее уменьшать большие коэффициенты.

Объяснение 3.

Проксимальный метод (позволяет эффективно минимизировать функции, в которых есть дифференцируемая часть и не дифференцируемая, но выпуклая).

$$Q(w) + \alpha||w||_1 \to \min$$

$$w^{(k)} = S_{\eta,\alpha}\left(w^{(k-1)} - \eta\nabla_w Q\big(w^{(k-1)}\big)\right)$$

$$S_{\eta,\alpha}(w_i) = \begin{cases} w_i - \eta\alpha, & w_i \ge \eta\alpha \\ 0, & |w_i| < \eta\alpha \\ w_i + \eta\alpha, & w_i < -\eta\alpha \end{cases}$$

Линейная классификация.

$\mathbb{Y} = \{1, \dots, k\}$ — многоклассовая классификация.

$\mathbb{Y} = \{-1, +1\}$ — бинарная классификация. (пока говорим про нее)

$\boxed{a(x) = sign < w, x >}$

Если $< w, x > = 0$:
1. такое событие невозможно.
2. отказ от классификации.
3. выдать случайный класс $(+1, -1)$

Геометрия:

$< w, x > = 0$ — уравнение гиперплоскости.
w — вектор нормали.
то есть линейный классификатор разделяет классы гиперплоскостью.
$|< w, x >|$ — тем больше, чем дальше x от гиперплоскости. Говорит об уверенности модели.

$$Q(a) = \frac{1}{\ell}\sum_{i=1}^{\ell}[a(x_i) \ne y_i] \text{ — доля ошибок } (error\ rate)$$

$$\frac{1}{\ell}\sum_{i=1}^{\ell}[sign < w, x_i > \ne y_i] \to \min_w,$$ к сожалению, за полиномиальное время не решить.

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\Big[\underbrace{y_i < w, x_i >}_{M_i:\ \text{отступ}} \ < \ 0\Big],$$ где $y_i = \pm 1$, так что если, $\begin{cases} y_i < w, x_i > \ > \ 0 & \Rightarrow \quad y_i = sign < w, x_i > \\ y_i < w, x_i > \ < \ 0 & \Rightarrow \quad y_i \ne sign < w, x_i > \end{cases}$

$signM_i$ — корректность классификации
$|M_i|$ — уверенность

То есть мы пока что используем следующую функцию потерь: $L(M) = [M < 0]$ - пороговая функция (фуфло)

Идея: $[M < 0] \le \tilde{L}(M)$ — дифференцируемая верхняя оценка.

$$\frac{1}{\ell}\sum_{i=1}^{\ell}[y_i < w, x_i > \ < \ 0] \quad \le \quad \frac{1}{\ell}\sum_{i=1}^{\ell}\tilde{L}(y_i < w, x_i >) \to \min$$

Hinge Loss: $\tilde{L}_1(M) = \max(0, 1 - M)$ Позволяет добиться дополнительной регуляризации.

Логистическая: $\tilde{L}_2(M) = \log(1 + \exp(-M))$ Позволяет оценивать вероятности

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\tilde{L}(y_i < w, x_i >) + \alpha R(w) \to \min_w$$ с помощью градиентного спуска.