


Problem 1

- Нет, не обязательно.
 - занимает веса именно L_1 регулар., а тут L_2 . Он их будет давать, но не занимать
 - даже с L_1 занятие не гарантировано в общем случае
 - но случайно может получиться и разреженный (переломить позыми?)
- Верно. Ошибка достигает минимума в задаче ее минимизации \equiv задачи без регуляризации $\equiv \lambda=0$. При росте λ мы будем сильнее штрафовать исходное решение и отходить от него.
 $w^* = (X^T X + \lambda I)^{-1} X^T y$
- Не верно. Может оказаться и так, что в общем случае, но регуляриз. борется с переобучением \Rightarrow скорее всего ошибка на тесте будет падать
 y SVM будет меньше смещение, чем y logit \Rightarrow т.е. он сконцентрирует на этом в обучении
 y logit будет меньше разброс
- Logit $\lambda=0$ vs SVM, Exp
 $\text{logit } L(M) = \ln(1 + \exp(-M))$
 $\text{SVM } L(M) = \max(0, 1-M)$

 - SVM не тратит ресурсы на увеличение отступа \rightarrow хорошо \Rightarrow просто разделяет.
 - Logit аккуратнее посчитает вероятности
- Нет, не стоит. Г.и. $\cos(x)$ - период. функция, то $\cos(\frac{\pi}{2} + 10^{10}) = \cos(\frac{\pi}{2}) = 0$ - так себе штраф за 10^{10} .

Problem 2

- Бэггинг оставляет смещение таким же, а разброс уменьшает (если модели не похожи друг на друга). В описанном способе будет куча неглубоких деревьев \rightarrow у них высокое смещение \rightarrow в среднем по композиции смещение будет больше, чем если бы все деревья были глубокими. Еще такая модификация снизит разброс, но не сильно \Rightarrow смысла в ней нет.
- Пусть $L(y, z)$ - дифф. функц. потерь. v_0 - начальная модель, например константа. v_N - итеративно, на шаге N

$$s_i = - \frac{\partial L(y_i, z)}{\partial z} \Big|_{z = a_{N-1}(x_i)}$$

$$v_N = \frac{1}{N} \sum_{i=1}^N (v_N(x_i) - s_i) \rightarrow \min_{v_N}$$

базовую v_N (составляющую общей модели) и s_i с помощью MSE

предыд. агрег. модель: $\hat{f}_N = \sum_{i=1}^N L(y_i, a_{N-1}(x_i) + v_N v_N(x_i)) \rightarrow \min_{\hat{f}_N \in R}$ - находит δ для модификации шага.

После этого общая модель на шаге N : $a_N(x) = \sum_{i=1}^N v_N(x)$
- RF - бэггинг. В нем смещение остается таким же, а разброс падает (см 2.1). Линейные модели же противоречат необходимости: - у них плохое смещение (не маленькое) \rightarrow RF будет с плохим смещ. - у них не выс. разброс \rightarrow после обучения мб станет меньше, если мы сможем сделать разные лин. модели.

Т.е. bias останется плохим, а разброс как был хорошим, так и останется. Ну и зачем? Мы же тогда не пользуемся плюсами RF.

В целом $L(\mu) = E_x E_y (y - \mu(X)(x))^2 = E_{xy} (y - E(y|x))^2 + E_x (E_x \mu(X) - E(y|x))^2 + E_x E_x (\mu(X)(x) - E_x \mu(X)(x))^2$
 Еще про 2 для обучения базовой модели мы используем именно MSE, а не L , т.к. L не релевантна к этой задаче, а MSE максимизирует угол наклона, чего мы и хотим.

Problem 3

Нет

Надо

RMS Prop

$$b_{k,i} = b_{k,i-1} + \frac{1}{2\eta} (\nabla_w Q(w^{k-1}))_i^2$$

$$b_{k,i} = b_{k,i-1} + (\nabla_w Q(w^{k-1}))_i^2$$

$$(или 2b_{k,i-1} + (1-2)(\nabla)_i^2)$$

$$w_j^k = w_j^{k-1} - \frac{\sqrt{b_{k,i} + \epsilon}}{\eta_k} (\nabla_w Q(w^{k-1}))_j$$

$$w_j^k = w_j^{k-1} - \frac{\eta_k}{\sqrt{b_{k,i} + \epsilon}} (\nabla_w Q(w^{k-1}))_j$$

• $\frac{1}{2\eta}$ в $b_{k,i}$ - лишнее. Оно приведет к быстрому затуханию обновления $b_{k,i}$, т.е. он перестанет расти как должен \rightarrow смысла всего Ada Grad пропадает, т.к. мы не начнем давать меньший вес тысячному обновлению (при $\frac{1}{\sqrt{b_{k,i} + \epsilon}}$), если до этого ходили много и далеко в этом направлении.

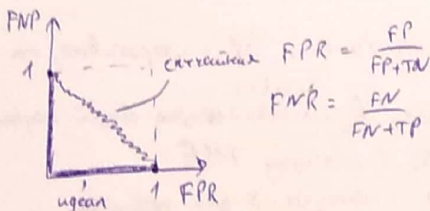
$$\frac{\sqrt{b_{k,i} + \epsilon}}{\eta_k} \rightarrow \frac{\eta_k}{\sqrt{b_{k,i} + \epsilon}}$$

мы увеличиваем вес вместо того, чтобы уменьшать !

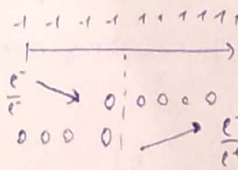
Но ведь весь Ada Grad нужен для того, чтобы отдельно по каждому признаку снижать размер следующих обновлений. А тут он растет \rightarrow алгоритм вообще может не сойтись, т.к. даже с шагом $\frac{1}{2\eta}$ мы выйдем на примерно постоянную длину шага.

А если мы еще и разделим на η_k (или η), которая скорее всего тоже падает, то длина шага будет расти, а алгоритм - расхожиться.

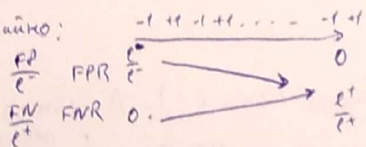
Problem 4



Для идеальной модели: кривая - угол. AUC = 1

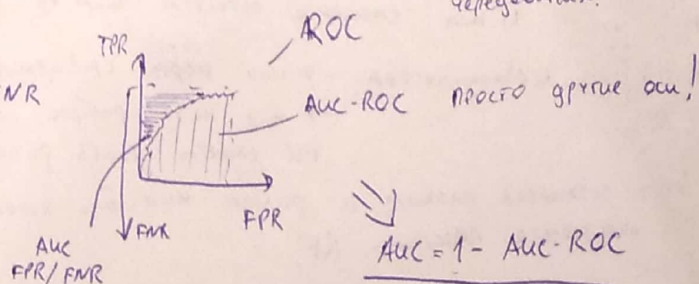


1. Случайно: $FPR = \frac{FP}{FP+TN}$, $FNR = \frac{FN}{FN+TP}$



т.е. они - диагональ $\Rightarrow AUC \approx \frac{1}{2}$ (не $\frac{1}{2}$ из-за возможного дисбаланса классов и разного возможного переувеличения).

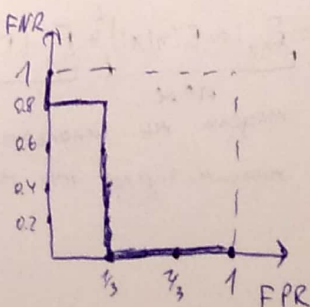
3. Заметим, что $TPR = \frac{TP}{TP+FN} = 1 - \frac{FN}{TP+FN} = 1 - FNR$. Значит, из ROC легко считать FNR/FPR



$$AUC = 1 - AUC-ROC$$

4. 0 0.2 0.3 0.4 0.5 0.7 0.9 0.95

	-1	-1	1	1	1	1	-1	1
TPR	$\frac{3}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0
FNR	0	0	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	$\frac{1}{5}$	$\frac{5}{5}$



$$AUC = \frac{1}{3} \cdot \frac{4}{5} = \frac{4}{15}$$

Проверим. AUC-ROC = площадь под кривой. Пар. Таких: $0.9 > \{0.3, 0.4, 0.5, 0.2\}$. Всего пар 15: $5! = 120$. Итого AUC ROC = $\frac{11}{15}$. $AUC = 1 - AUC-ROC = 1 - \frac{11}{15} = \frac{4}{15}$.