1. Докажите, что задача логистической регрессии является выпуклой задачей машинного обучения (строго обоснуйте каждый переход, например, если говорите, что выпукла, так как композиция линейной функции и выпуклой, то обозначьте, какая функция линейна, а какая выпуклая)

Множество гипотез

$$H_{\mathrm{sig}} = \phi_{\mathrm{sig}} \circ L_d = \{\mathbf{x} \mapsto \phi_{\mathrm{sig}}(\langle \mathbf{w}, \mathbf{x} \rangle) : \mathbf{w} \in \mathbb{R}^d\}.$$

будет являться выпуклым т.к. $H = \mathbb{R}^d$

Сама задача имеет вид

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 + \exp(-y_i \langle \mathbf{w}, \mathbf{x}_i \rangle)\right).$$

Доменом будет $Z = X \times Y = \mathbb{R}^{d+1}$

Рассмотрим ф-ию потерь на этом домене.

$$\ell(h_{\mathbf{w}}, (\mathbf{x}, y)) = \log\left(1 + \exp(-y\langle \mathbf{w}, \mathbf{x} \rangle)\right).$$

И покажем, что она выпукла.

Рассмотрим ф-ию g(x) = log (1 + exp(x)). И найдем её вторую производную.

g'(x) = $\frac{exp(x)}{1+exp(x)}$

g''(x) = $\frac{exp(x)*(1+exp(x)) - exp(x)*exp(x)}{(1+exp(x))(1+exp(x))}$ = $\frac{exp(x)}{(1+exp(x))^2}$ >= 0

Следовательно g(x) является выпуклой.

Ф-ия f(x) = -y<w,x> является линейной.

Тогда фи-ия потерь будет выпуклой как композиция выпуклой ф-ии f(x) и линейной ф-ии g(x).

Следовательно, задача логистической регрессии является выпуклой задачей машинного обучения.

3. Докажите, что функция является выпуклой тогда и только тогда, когда её надграфик является выпуклым множеством

Пусть f(x) выпуклая ф-ия. f: $R^n \to R$

Покажем, что её надграфик выпуклый.

Пусть (x1,y1) и (x2,y2) точки надграфика f(x). Рассмотрим линию вида (x’,y’)=a(x1,y1)+(1-a)(x2,y2) , где a ∈ $[0, 1]$
Тогда
y’ = ay1+(1-a)y2 >= [т.к. это точки надграфика, то y1>=f(x1)]
>=af(x1)+(1-a)f(x2)>=f(ax1+(1-a)x2)=f(x’)
Т.е. y’ лежит выше чем f(x’),значит, он содержится в надграфике (все точки линии лежат в надграфике). Следовательно надграфик это выпуклое множество.

Теперь в другую сторону.
Пусть надграфик выпуклый и x1, x2 из R^n ,a ∈ $[0, 1]$
Тогда (x1, f(x1)) и (x2,f(x2)) находятся в надграфике f(x), а так как он выпуклый то тогда a(x1,f(x1))+(1-a)(x2,f(x2)) тоже лежит в надграфике. Тогда f(a*x1+(1-a)x2) <= af(x1) + (1-a)f(x2). Значит, f(x) выпуклая.

2. Докажите, что в случае разделимой выборки задача оптимизации функционала логистической регрессии не имеет решения, если не использовать регуляризацию

Функционал функции логистической регрессии имеет вид

$$\sum_{i=1}^{\ell} \ln\bigl(1 + \exp\left(-y_i \langle \boldsymbol{w}, x_i \rangle\right)\bigr) \to \min_{\boldsymbol{w}}.$$

Так как выборка разделима, то тогда существует вектор w такой что:
<w,xi> > 0 если yi = 1 и <w,xi> < 0 если yi = -1. Это и будет вектор разделяющий выборку (он может быть и не один). Тогда можно зафиксировать такой вектор и увеличивать его норму (домножая каждый элемент вектора на константу) и новый вектор также будет разделяющим, но при этом значение функционала можно будет постоянно уменьшать (оно будет как бы стремиться к нулю, но нуля мы достичь не сможем).
Т.е. в случае разделимой выборки у нас будет бесконечно много вариантов выбора вектора w. И решения задачи не будет существовать

А если будет регуляризация то выберем вектор с меньшей нормой.