CS534 Machine learning EX2

Kaiwen Zheng

**0. List your team members for HW1.**

Haoyu Zhang, Xiao Tan

**1. State your individual contributions to your group's HW1 submission.**

Discuss with teammate with the model, do the code in the HW.

**3. Some groups included features that are not observed on the training data but**

**might appear on dev/test, such as age=0, age=1, ..., age=100, etc., while other**

**groups simply used observed features from the training set.**

**Which option is better, or is there any difference?**

In my opinion, appear more could be better it might exit some features that would not appear in the training set. Even if the new data don’t have the new features, the model would not observer these features change so that it would not have influence on result. Also it might need more time to generate the model.

**4. You have observed that machine learning tends to exaggerate the existing bias in data.**

**For example, there are about 24% positive examples in the dev set, but your models**

**most likely predicted only around 20% positive on that set.**

**(a) is this due to overfitting or underfitting?**

**(b) is the exaggeration more severe on training set or on dev set? why?**

**(c) if we don't observe any feature, is it better to predict 100% or 75% negative? why?**

**(d) what other biases can you find that your best HW1 model exaggerates on the dev set?**

**e.g., what about people with a Doctoral degree?**

(a) It would be underfitting

(b) The exaggeration more severe more on training set.

(c) Predict 100% negative. Because if we just predict negative the result would be the p (the negative percentage on the total). And if predict 75% negative, it would p\*75%+(1-p)\*25%=0.25+0.5p, because p is 75% in this test so it would 75% negative would be smaller than the 100% negative prediction.

(d) In the model build by us, the bias(b) would be -5 and the degree part such a such Doctoral degree would be 64, therefore the weight of this would be bias.

**5. For your best model,**

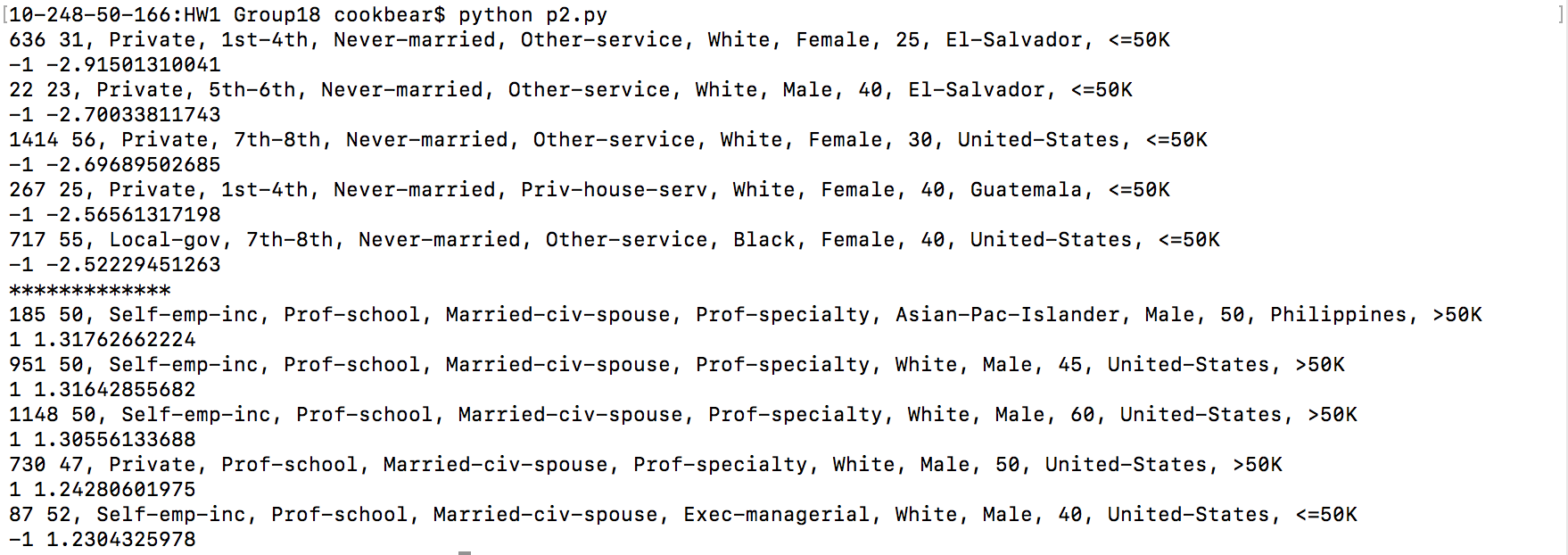
**(a) what are the top five most positive/negative examples on the dev set?**

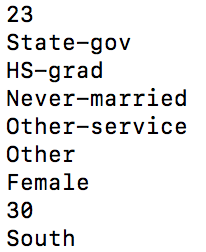
**(b) did your model predict them correctly?**

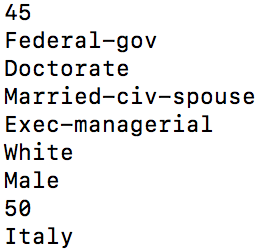
**(c) for each column (e.g., age), what is the most positive/negative feature?**

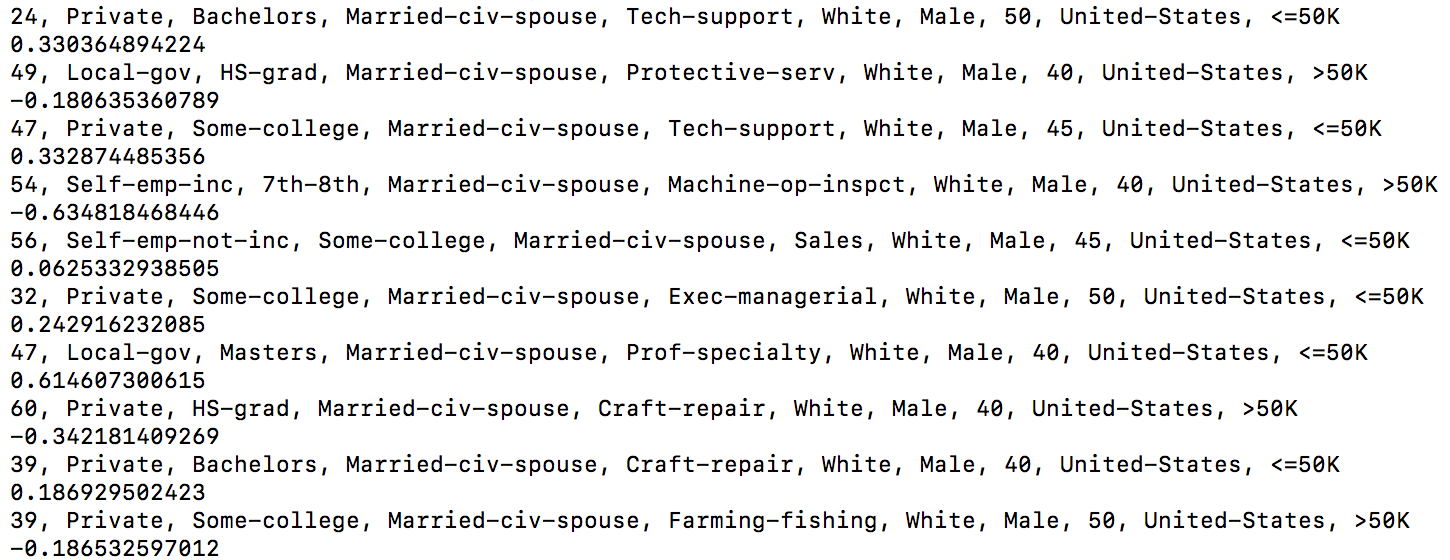
**(d) list five incorrectly predicted examples on the dev set. any observations?**

(a) and (b) just the last one is incorrect

****

(c) the most negative in each column

the most positive in each column

(d) 

So feature would make some special example.

**Part II: SVM Theory**

**-------------------**

**1. State the two equivalent formulations for SVMs:**

**(a) maximize geometric margin**

**(b) minimize weight norm**

(a)

(b)

**2. What will happen**

**(a) if we replace the minimum functional margin of 1 by 10**

**(b) if we require the geometric margin to be at least 1**

(a) Because we get the and when replace 1 by 10, we must get the new W which is 10 times of original w.

(b) It would depend on the data set. For the data set, we may get different result w. If the bigger than 1 it would make scale to 1. If the smaller than 1, it would be not impossible to scale the w.

**3. How many support vectors can there be for SVMs in d dimensions? (start with d=1,2,3)**

It would depend on the dimensions, d or d+1, but some time would be d+2 when the data set is not separable.

**4. True or False?**

**a) if an example has functional margin of 1, it must be in the support vector set.**

**b) if an example is in the support vector set, it must have a functional margin of 1.**

(A) F (B) T

**5. List two reasons why the convex hull approach is not used to solve SVMs in practice.**

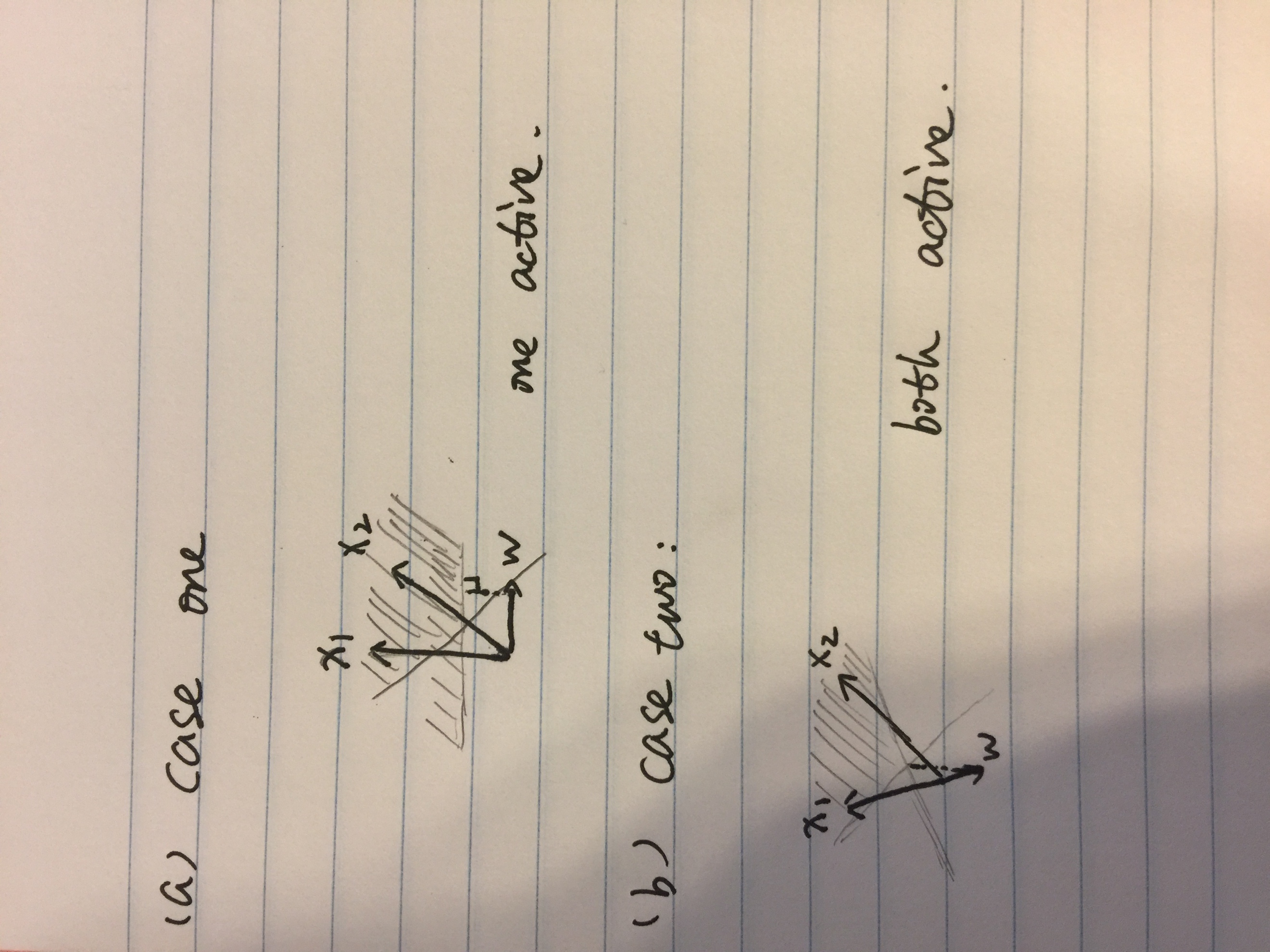
(1) Because when the feature dimension grows bigger, the convex hull approach would be much slower in the practice.

(2) Because the convex hull would calculate much more than what we want in the model about the support vector, then it would may be slower.

**6. Draw an example update of minibatch MIRA where the batch size is 2, i.e., you consider two examples at a time, and after the update, your new model should achieve a functional margin of at least 1 on both examples. Assume both examples are incorrectly predicted by the current model, and one example is positive while the other is negative.**

**(a) case one: one constraint is active.**

**(b) case two: both constraints are active.**

****

**7. For a separable dataset, how many support vectors can there be for perceptron?**

A: According to the number of feature, basically is 2 support vectors in the 2d dimension.

**8. Why convex optimization is (much) easier than general optimization?**

In general, the convex optimization is based on the convex set, general optimization is based on the convex function. Also, because the convex function is over the convex set, therefore, convex optimization would be a subset of the general optimization.

**9. Are all quadratic programs instances of convex optimization?**

**Are all linear programs instances of convex optimization?**

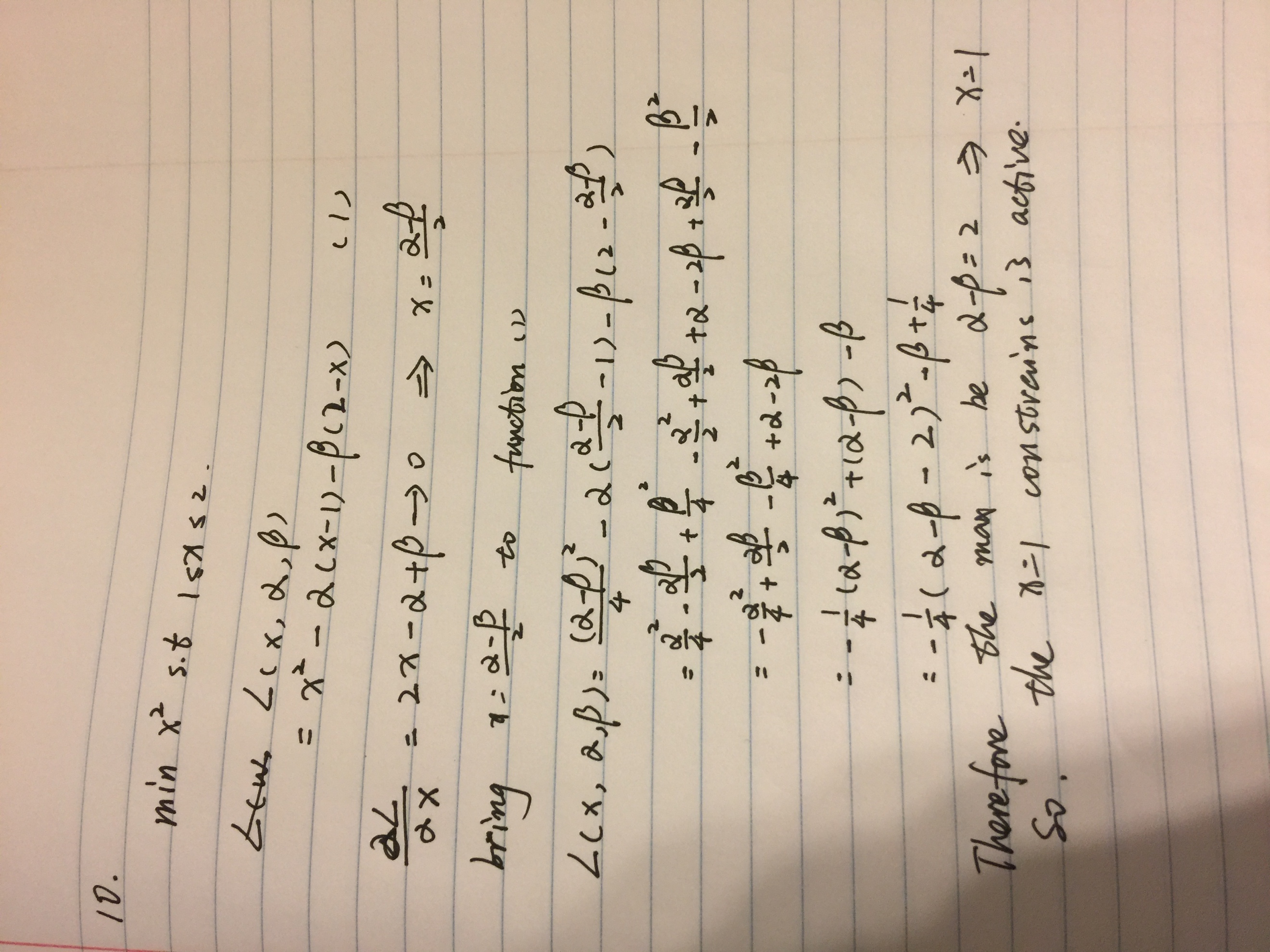
**Why SVM is convex optimization?**

No all quadratic programs instances are convex optimization. Just convex QP is the convex optimization. QP is NP-hard problem to solved.

All linear programs instances are convex optimization. Because we use the KKT to satisfied the SVM model with the margin get the 1. Therefore, we must get the activing constrain conditions at most 2 rather than the all conditions. Then the SVM is the subset the convex hull problem.

**10. Use Lagrangian multipliers to solve: min x^2, where 1<=x<=2.**

**Which constraints are active?**

****

**11. Explain the "complementary slackness" condition in KKT in your own words.**

Because the KKT condition require the active condition which the margin should be 1. Also, the problem can be defined to . The loss function would be .The optimal point would have satisfied at most two conditions and other condition/loss function would be positive, which would be part would make the margin slack . would be each vector slack. Because the loss function would not be always 0, the vector would slack at most time.

**12. (optional) Why the alphas in Lagrange multipliers for inequality constraints must be non-negative?**

We want to find the result and .(x>=b) We can find different case, when x = b, can be anything. When x>b, would be find result when . If a is negative, it would be , it can’t find the result. Consider the upper reason, the should not be negative.

**13. (optional) Why minimizing the original constrained optimization objective becomes maximizing the new objective on alphas? (the optimal answer is achieved on a saddle point)**

Due to we would solve the problem from primal to dual, we would change the minimize function to maximize function subject to . It would