FINAL EXAM, Jan 16th, 2017

Machine Learning Course

Fall 2016

EPFL

School of Computer and Communication Sciences

Martin Jaggi & Rüdiger Urbanke

mlo.epfl.ch/page-136795.html

epfmlcourse@gmail.com

Rules:

- This exam counts 60 percent towards your final grade.
- You have 180 minutes (from 16:15 till 19:15) to complete the exam.
- This is a closed book exam.
- No electronic devices of any form (mobile, calculator, ...) are allowed.
- Place on your desk: your student ID, writing utensiles, one double-sided handwritten A4 page of summary if you have one, highly sugary drinks and ... if you need them.
- Place all your other personal belongings at the entrance or under your desk
- Write all your answers on the provided space. If you need extra paper let us know.

Good Luck!

Blue Exam Version

Prob I	/ 57
Prob II	/ 5
Prob III	/ 10
Prob IV	/ 10
Prob V	/ 20
Total	/ 102

1 Multiple Choice Questions and Simple Problems [57pts]

Mark the correct **answer(s)**. More than one answer can be correct!

1. (2 pts) Let the samples $\{(y_n, x_n)\}$ come from some fixed joint distribution $p(x, y)$, where x_n and y_n are scalars and both have zero mean. Consider *linear* regression, i.e., we want to predict Y from X by means of $f(x) = \alpha x$ and we consider a square loss. Meaningful regression is possible

   (a) □ only if X "causes" Y

   (b) □ as long as Y and X have non-zero correlation

   (c) □ only if Y and X are positively correlated, i.e., $\mathbb{E}[XY] > 0$

   (d) □ only if Y and X are negatively correlated, i.e., $\mathbb{E}[XY] < 0$

   Solution: Answer b is correct. In fact a quick calculations shows that the optimum choice is $\alpha = \frac{\mathbb{E}[XY]}{\mathbb{E}[X^2]}$.

2. (2 pts) Consider a linear regression problem with N samples where the input is in D-dimensional space, and all output values are $y_i \in \{-1, +1\}$. Which of the following statements is correct?

   (a) □ linear regression cannot "work" if $N \gg D$

   (b) □ linear regression cannot "work" if $N \ll D$

   (c) □ linear regression can be made to work perfectly if the data is linearly separable

   Solution: Answer c is correct.

3. (2 pts) You have data with lots of outliers. Everything else being equal, and assuming that you do not do any pre-processing, what cost function will be less effected by these outliers?

   (a) □ $(y - f(x))^2$ (MSE)

   (b) □ $|y - f(x)|$ (MAE)

   Solution: Answer b is correct since it weighs large outliers less heavily.

4. (4 pts) The following function(s) have a unique minimizer.

   (a) □ $f(x) = x^2$, $x \in [-3, 2]$

   (b) □ $f(x) = \log(x)$, $x \in (0, 10]$

   (c) □ $f(x) = \sin(x)$, $x \in [-10, 10]$

   (d) □ $f(x) = e^{3x} + x^4 - 3x$, $x \in [-10, 10]$

   Solution: The examples in a and d are both strictly convex and hence have a unique minimizer. The logarithm does not take on its minimum in $(0, 10)$ and the sinusoid has multiple minima.

5. (2 pts) What is the gradient of $\mathbf{x}^\top \mathbf{W} \mathbf{x}$ with respect to all entries of $\mathbf{W}$ (written as a matrix)?

   (a) □ $\mathbf{W}\mathbf{x}$.

   (b) □ $\mathbf{W}^\top \mathbf{x}$.

   (c) □ $(\mathbf{W} + \mathbf{W}^\top)\mathbf{x}$.

   (d) □ $\mathbf{W}$

   (e) □ $\mathbf{x}\mathbf{x}^\top$.

   (f) □ $\mathbf{x}^\top \mathbf{x}$

   (g) □ $\mathbf{W}\mathbf{W}^\top$.

   Solution: Answer e is correct. Writing out the expression explicitly we have $\sum_{i,j} W_{i,j} \mathbf{x}_i \mathbf{x}_j$. Taking the derivative with respect to $\mathbf{W}_{i,j}$ we hence get $\mathbf{x}_i \mathbf{x}_j$.

6. (2 pts) What is the gradient of $\mathbf{x}^\top \mathbf{W} \mathbf{x}$ with respect to $\mathbf{x}$ (written as a vector)?

   (a) □ $\mathbf{W}\mathbf{x}$.

   (b) □ $\mathbf{W}^\top \mathbf{x}$.

(c) ☐ $(\mathbf{W} + \mathbf{W}^\top)\mathbf{x}$.

(d) ☐ $\mathbf{W}$

(e) ☐ $\mathbf{x}\mathbf{x}^\top$.

(f) ☐ $\mathbf{x}^\top\mathbf{x}$

(g) ☐ $\mathbf{W}\mathbf{W}^\top$.

Solution: Answer c is correct. Writing out the expression explicitly we have $\sum_{i,j} W_{i,j}\mathbf{x}_i\mathbf{x}_j$. Taking the derivative with respect to x_i we get two terms, once corresponding to $\mathbf{W}\mathbf{x}$ and one corresponding to $\mathbf{x}^\top\mathbf{W}$.

7. (2 pts) The following member of the exponential family represents a scalar Gaussian: $p(y) = \exp\{(2, -1)(y, y^2)^\top - 1 - \frac{1}{2}\ln(\pi)\}$. What are the mean μ and the variance σ^2?

   (a) ☐ $\mu = -1$, $\sigma^2 = 0$.

   (b) ☐ $\mu = 0$, $\sigma^2 = 0$.

   (c) ☐ $\mu = 1$, $\sigma^2 = 0$.

   (d) ☐ $\mu = -1$, $\sigma^2 = \frac{1}{2}$.

   (e) ☐ $\mu = 0$, $\sigma^2 = \frac{1}{2}$.

   (f) ☐ $\mu = 1$, $\sigma^2 = \frac{1}{2}$.

   (g) ☐ $\mu = -1$, $\sigma^2 = 1$.

   (h) ☐ $\mu = 0$, $\sigma^2 = 1$.

   (i) ☐ $\mu = 1$, $\sigma^2 = 1$.

   Solution: Answer f is correct.

8. (2 pts) Consider a learning algorithm that has the property that it depends only very weakly on the input data. E.g., this could be SGD where we choose a very small step size and only run for very few iterations. To go to the extreme, you can imagine a learning algorithm that always outputs the same model irrespective of the training set. Presumably such a learning algorithm will not give us good results. Why is that?

   (a) ☐ Such a learning algorithm typically has a much larger generalization error than training error.

   (b) ☐ Such a learning algorithm typically has a large bias.

   (c) ☐ Such a learning algorithm is prone to overfitting.

   Solution: Answer b is correct. Since the model only weakly depends on the training data it is unlikely going to model it well.

9. (2 pts) You have given the 2D data shown in Figure 1. You are allowed to add one component to your data (in addition to a constant component) and then must use a linear classifier. What component should you pick? [*HINT:* The axes are not labeled so you cannot check it mathematically, but the shape of the data is very suggestive and exactly one answer is correct.]

   (a) ☐ $1/|x_1|$

   (b) ☐ $1/|x_2|$

   (c) ☐ $|x_1 x_2|$

   (d) ☐ $\log_2 |x_1 + x_2|$

   (e) ☐ $x_1^2 + x_2^2$

   (f) ☐ $1/|x_1 + x_2|$

   Solution: Answer c is correct. The figure suggests that the labels have the form $\mathbf{1}_{\{|x_1 x_2| \leq c\}}$. Hence, this is the component we should add.

10. (2 pts) Assume you run k-nearest neighbor given the data set of the previous problem (Figure 1). Mark the correct statements.

    (a) ☐ k-nearest neighbors with $k = 1$ would work well.

    (b) ☐ k-nearest neighbors with $k = 1$ would work better if we use it on the "extended" data with one component added as discussed in the previous problem.

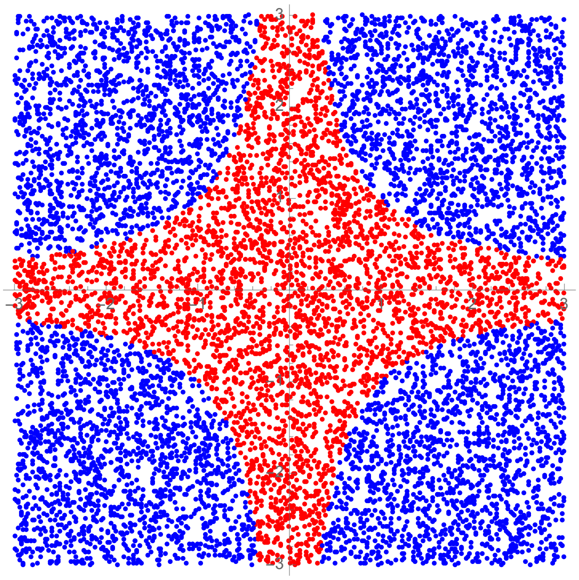

Figure 1: Some 2D data for classification.

(c) ☐ k-nearest neighbors with $k = 1$ only works on data that is linearly separable.

Solution: Answer a is correct. We have lots of data, work in small dimensions and the decision boundaries are fairly simple. Adding one extra dimension does not help at all for the k-nearest neighbor and in fact likely will make performance worse since we are operating now in higher dimension. The last assertion is also wrong. Linear separability helps greatly for linear classifiers but has little to do with nearest neighbor classifiers.

11. (2 pts) Consider the K-means algorithm. Which of the following assertions is correct?

(a) ☐ Regardless of the initialization the algorithm converges.

(b) ☐ Regardless of the initialization the algorithm always finds the same clusters.

(c) ☐ If we initialize the K-means algorithm with optimal clusters then it will find in one step optimal representation points.

(d) ☐ If we initialize the K-means algorithm with optimal representation points then it will find in one step optimal clusters.

Solution: Answers a, c, and d are correct.

12. (2 pts) Mark the true statements.

(a) ☐ Logistic loss is typically preferred over L_2 loss (least squares loss) in classification tasks.

(b) ☐ In terms of feature selection, L_2 regularization is often preferred since it comes up with sparse solutions.

Solution: Answer a is correct. With logistic loss, correctly classified points that are far away from the decision boundary have much less impact on the decision boundary. W.r.t. the second question, L_1 regularization (LASSO) comes up with sparse solutions.

13. (2 pts) K-means can be equivalently written as the following Matrix Factorization

$$\min_{\mathbf{z},\boldsymbol{\mu}} \ \mathcal{L}(\mathbf{z},\boldsymbol{\mu}) = \|\mathbf{X} - \mathbf{M}\mathbf{Z}^\top\|^2_{\text{Frob}}$$
$$\text{s.t. } \boldsymbol{\mu}_k \in \mathbb{R}^D,$$
$$z_{nk} \in \mathbb{R}, \sum_{k=1}^{K} z_{nk} = 1.$$

(a) ☐ yes

(b) ☐ no

Solution: No, z must be integer

14. (2 pts) In a Gaussian Mixture Model, assuming $D, K \ll N$, the number of free parameters, after marginalization of the latent variables z_n, is

    (a) ☐ quadratic in D

    (b) ☐ cubic in D

    (c) ☐ linear in N

    Solution: a is correct; in fact the number is $O(D^2K)$

15. (2 pts) In the setting of EM, where x_n is the data and z_n is the latent variable, what quantity is called the posterior?

    (a) ☐ $p(\mathbf{x}_n|z_n, \boldsymbol{\theta})$

    (b) ☐ $p(\mathbf{x}_n, z_n|\boldsymbol{\theta})$

    (c) ☐ $p(z_n|\mathbf{x}_n, \boldsymbol{\theta})$

    Solution: c) is correct

16. (2 pts) Matrix Factorizations: The function $f(\mathbf{v}) := g(\mathbf{v}\mathbf{v}^\top)$ is convex over the vectors $\mathbf{v} \in \mathbb{R}^2$, when $g : \mathbb{R}^{2\times 2} \to \mathbb{R}$ is defined as

    (a) ☐ if we define $g : \mathbb{R}^{2\times 2} \to \mathbb{R}$ as $g(\mathbf{X}) := X_{11}$.

    (b) ☐ if we define $g : \mathbb{R}^{2\times 2} \to \mathbb{R}$ as $g(\mathbf{X}) := X_{11} + X_{22}$.

    Solution: Both are correct.

17. (2 pts) Matrix Factorizations: If we compare SGD vs ALS for optimizing a matrix factorization of a $D \times N$ matrix, for large D, N

    (a) ☐ Per iteration, SGD has a similar computational cost as ALS

    (b) ☐ Per iteration, ALS has an increased computational cost over SGD

    (c) ☐ Per iteration, SGD cost is independent of D, N

    Solution: b) and c) are correct

18. (2 pts) Text:

    (a) ☐ Comparing the word feature representations from bag-of-words vs GloVe, bag-of-words typically gives lower dimensional representations.

    (b) ☐ GloVe and word2vec are typically trained unsupervised

    Solution: b) is correct

19. (2 pts) Assume that we have a data matrix $\mathbf{X}$ of dimension $D \times N$ as usual. Suppose that its SVD is of the from $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$, where $\mathbf{S}$ is a diagonal matrix with $s_1 = N$ and $s_2 = s_3 = \cdots = s_D = 1$. Assume that we want to compress the data from D to 1 dimensions via a linear transform represented by a $1 \times D$ matrix $\mathbf{C}$ and reconstruct then via $D \times 1$ matrix R. Let $\hat{\mathbf{X}} = \mathbf{R}\mathbf{C}\mathbf{X}$ be the reconstruction. What is the smallest value we can achieve for $\|\mathbf{X} - \hat{\mathbf{X}}\|_F^2$?

    (a) ☐ D

    (b) ☐ $D - 1$

    (c) ☐ $N - D$

    (d) ☐ $N - D + 1$

    (e) ☐ $N - D - 1$

    (f) ☐ $N - 1$

    (g) ☐ N

    (h) ☐ ND

Solution: Answers b is correct. We have learned that the Frobenius norm of the difference is at least equal to the sum of squares of the singular values that we leave out. In our case each of them has value 1 and we leave out $D-1$ of them.

20. (2 pts) Which of the following statements is correct?

    (a) ☐ A neural net with one hidden layer and an arbitrary number of hidden nodes with sigmoid activation functions can approximate any "suffiently smooth" function.

    (b) ☐ A neural net with one hidden layer and an arbitrary number of hidden nodes with sigmoid activation functions can approximate any "suffiently smooth" function on a bounded domain.

    (c) ☐ On a bounded domain, neural nets can approximate any "sufficiently smooth" function "in average" but not "pointwise".

    Solution: Answer c is correct. We can only approximate functions in bounded domains. But we can approximate them in average or pointwise.

21. (2 pts) The complexity of the back-propagation algorithm for a neural net with L layers and K nodes per layer is

    (a) ☐ $\Theta(K^L)$

    (b) ☐ $\Theta(L^K)$

    (c) ☐ $\Theta(K^2L^2)$

    (d) ☐ $\Theta(K^2L)$

    (e) ☐ $\Theta(KL^2)$

    (f) ☐ $\Theta(KL)$

    (g) ☐ $\Theta(K)$

    (h) ☐ $\Theta(L)$

    (i) ☐ $\Theta(1)$

    Solution: Answer d is correct. The dominant term is a multiplication of a vector with a $K \times K$ matrix and this has to be done in each of the L layers.

22. (2 pts) Assume that you initialize all weights in a neural net to the same value and you do the same for the bias terms. Which of the following statements is correct.

    (a) ☐ This is a good idea since it treats every edge equally.

    (b) ☐ This is a bad idea.

    Solution: Answer b is correct since in this case every node on a particular level will learn the same feature.

23. (2 pts) [Gradient for convolutional neural nets] Let $f(x,y,z,u,v,w) = 3xyzuvw + x^2y^2w^2 - 7xz^5 + 3yvw^4$. What is

$$\left[\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} + \frac{\partial f}{\partial z} + \frac{\partial f}{\partial u} + \frac{\partial f}{\partial v} + \frac{\partial f}{\partial w}\right]\Big|_{x=y=z=u=v=w=1} ?$$

    (a) ☐ -4

    (b) ☐ -3

    (c) ☐ -2

    (d) ☐ -1

    (e) ☐ 0

    (f) ☐ 1

    (g) ☐ 2

    (h) ☐ 3

    (i) ☐ 4

    Solution: Answer e is correct since it is equal to $\frac{d}{dx}f(x,x,x,x,x,x)|_{x=1} = \frac{d}{dx}0 = 0$.

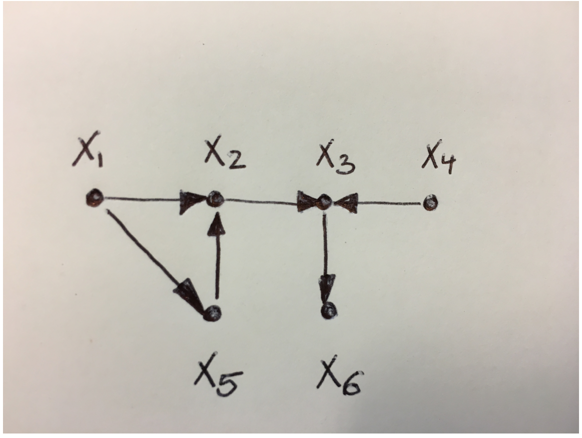

Figure 2: A Bayes net with six nodes.

24. (2 pts) Consider a neural net with K nodes per hidden layer. In a standard such net we have K^2 parameters/weights (ignoring the bias terms) per layer. Consider a convolutional net where the data is laid out in a one-dimensional fashion and the filter/kernel has M non-zero terms. Ignoring the bias terms, how many parameters are there per layer?

    (a) ☐ K^2
    (b) ☐ M^2
    (c) ☐ KM
    (d) ☐ K
    (e) ☐ M
    (f) ☐ 1

    Solution: Answer e is correct since we use weight sharing there are only M parameters.

25. (5 pts) Consider the Bayes Net shown in Figure 2. Which of the following statements is true?

    (a) ☐ X_1 and X_4 are independent.
    (b) ☐ X_1 and X_4 are independent given X_6.
    (c) ☐ X_1 and X_4 are independent given X_2.
    (d) ☐ X_1 and X_4 are independent given X_2 and X_3.
    (e) ☐ X_1 and X_4 are independent given X_5.

    Solution: Answers a, c, d, e are all correct. (a) Yes, since X_3 blocks all paths. (b) No, since X_3 no longer blocks the paths. (c) Yes, since X_2 blocks all paths. (d) Yes, since X_2 blocks all paths. (e) Yes, since X_3 blocks all paths.

26. (2 pts) Consider the factor graph shown in Figure 3. Assume that all random variables take values in the finite (but large) alphabet $\mathcal{X}$ of size $|\mathcal{X}|$. Note that this graph is a tree so that we can use the sum-product algorithm to compute the marginals. The complexity of computing the marginals (measured in function evaluations and simple operations such as additions and multiplications) is

    (a) ☐ $\Theta(|\mathcal{X}|^1)$
    (b) ☐ $\Theta(|\mathcal{X}|^2)$
    (c) ☐ $\Theta(|\mathcal{X}|^3)$
    (d) ☐ $\Theta(|\mathcal{X}|^4)$
    (e) ☐ $\Theta(|\mathcal{X}|^5)$
    (f) ☐ $\Theta(|\mathcal{X}|^6)$

    Solution: Answers c is correct since the largest function degree is 3.

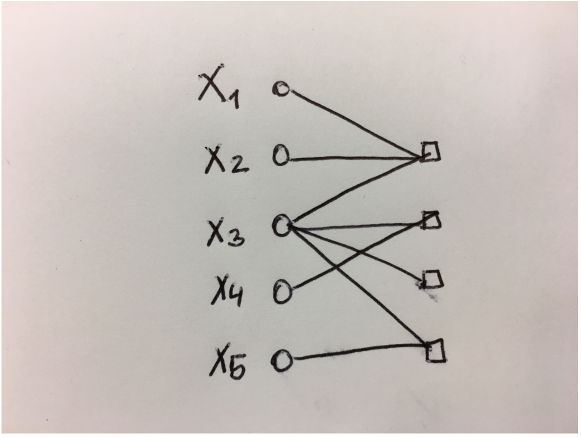

Figure 3: A Factor Graph.

.

2 Subgradients [5pts]

Compute a subgradient for the mean average error (MAE) cost function. That is for any $\mathbf{w}$, give a subgradient $\mathbf{g}$ at $\mathbf{w}$. [Hint: you are allowed to assume the gradient of f is known at any $\mathbf{w}$.]

$$\mathcal{L}(\mathbf{w}) = \mathsf{MAE}(\mathbf{w}) := \frac{1}{N}\sum_{n=1}^{N}\left|y_n - f(\mathbf{w}, \mathbf{x}_n)\right|$$

Solution: Subgradients obey the chain rule just like regular derivatives. Let $g(\mathbf{w}) = \nabla\mathcal{L}(\mathbf{w})$. Further, let

$$g(x) = \begin{cases} 1, & x > 0, \\ -1, & x < 0, \\ 0, & x = 0. \end{cases}$$

Then, using the chain rule we have

$$g(\mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N} h(y_n - f(\mathbf{w}, \mathbf{x}_n))\nabla y_n - f(\mathbf{w}, \mathbf{x}_n).$$

(scratch space)

3 Linear regression [10pts]

Consider a data matrix $\mathbf{X} \in \mathbb{R}^{D\times N}$ of N data points in D dimensions, and target values y_n for $n = 1, \cdots, N$. We perform least squares linear regression, without the use of any regularizer.

1. (5pts) Write down the normal equations.
2. (5pts) Give the expression to predict a new unseen point $\mathbf{x}_m$. Do not assume knowledge of $\mathbf{w}^\star$ but compute it.

(scratch space)

4 K-Means Clustering [10pts]

Consider a data matrix $\mathbf{X} \in \mathbb{R}^{D \times N}$ of N data points in D dimensions. We want to perform the K-means algorithm with the Euclidean distance on $\mathbf{X}$ with the addition that we also want to minimize the ℓ_2-norm of each cluster center $\mathbf{u}_k$.

$$\min_{\mathbf{U},\mathbf{Z}} \left[J(\mathbf{U}, \mathbf{Z}) := \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} z_{kn} \|\mathbf{x}_n - \mathbf{u}_k\|_2^2 + \frac{1}{2} \sum_{k=1}^{K} \|\mathbf{u}_k\|^2 \right]$$

Derive the update rule for the cluster centers $\mathbf{u}_k$.

(scratch space)

5 Matrix Factorization [20pts]

(Problem due to Alex Smola) Consider the following matrix-factorization problem. For the observed ratings r_{um} for a given pair (u, m) of a user u and a movie m, one typically tries to estimate the score by

$$f_{um} = \langle \mathbf{v}_u, \mathbf{w}_m \rangle + b_u + b_m.$$

Here $\mathbf{v}_u$ and $\mathbf{w}_m$ are vectors in $\mathbb{R}^D$ and b_u and b_m are scalars, indicating the bias.

1. (5 pts) Assume that our objective is given by

$$\frac{1}{2}\sum_{u\sim m}(f_{um} - r_{um})^2 + \frac{\lambda}{2}\left[\sum_{u\in\mathbf{U}}(b_u^2 + \|\mathbf{v}_u\|^2) + \sum_{m\in\mathbf{M}}(b_m^2 + \|\mathbf{w}_m\|^2)\right]$$

   where $\lambda > 0$. Here $\mathbf{U}$ denotes the set of all users, M the set of all movies, and $u \sim m$ represents the sum over all (u, m) pairs for which a rating exists. Write the optimal values of b_u, provided that all other values are fixed.

   Solution: Let us compute the derivative wrt a particular user u' and set it to 0. We get

$$\sum_{u'\sim m}(f_{u'm} - r_{u'm}) + \lambda b_{u'} = 0.$$

   Note that the $f_{u'm}$ contains the $b_{u'}$. Solving this equation for $b_{u'}$ we get

$$b_{u'} = \frac{\sum_{u'\sim m}(r_{u'm} - \langle \mathbf{v}_{u'}, \mathbf{w}_m \rangle - b_m)}{\lambda + \sum_{u'\sim m} 1},$$

   where $u' \sim m$ are the movies rated by u'.

2. (10 pts) Is the problem jointly convex in $\mathbf{v}$ and $\mathbf{w}$? Look at a simple case, say for only 1 user and 1 movie and assume that $D = 1$, i.e., consider $f(v, w) = \frac{1}{2}(vw + c - r)^2$. [Hint: A 2×2 matrix is positive definite if and only if the two diagonal terms are positive and the determinant is positive.]

   Solution: The Hessian of $f(v, w)$ is equal to

$$\begin{pmatrix} 2w^2 & 4vw - 2r \\ 4vw - 2r & 2v^2 \end{pmatrix}.$$

   This matrix is not positive semi-definite in general. Its determinant is equal to $-4(r - 3vw)(r - vw)$ which is not necessarily positive. Therefore the problem is only element-wise convex but not jointly convex in v and w.

3. (2.5 pts) How could you address the problem of recommending movies to a new user without any ratings? [This is not a math question.]

   Solution: When a new user come in, we have little information about them, and thus the matrix factorization method can not learn much associations between the new user and the existing users. We should use the demographics information of the user to bridge its associations with existing users.

4. (2.5 pts) How could you address the problem of potentially recommending a new movie without any ratings to users? [As in the previous point, this is also not a math question.]

   Solution: Using meta data of the movie as additional information to encode the similarity, perhaps approximating the corresponding weight as a linear combination of existing movies based on their similarities in terms of meta information.

(scratch space)

(scratch space)

(scratch space)

(scratch space)