Learning From Data
Caltech - **edX** CS1156x
courses.edx.org/courses/CaltechX/CS_1156x/3T2014/info
Fall 2014

Final Exam

Due **Friday**, December 5, 2014, at 22:00 GMT/UTC

All questions have multiple-choice answers ([**a**], [**b**], [**c**], **...**). *You can collaborate with others, but do not discuss the selected or excluded choices in the answers. You can consult books and notes, but not homework solutions. Your solutions should be based on your own work. Definitions and notation follow the lectures.*

Note about the final

- There are twice as many problems in this final as there are in a homework, and some problems require packages that will need time to get to work properly.
- More time is given to solve the final exam compared to the homework. Please note the above-mentioned due date (due Friday of next week, not Monday).
- Problems cover different parts of the course. To facilitate your search for relevant lecture parts, an indexed version of the lecture video segments can be found at the Machine Learning Video Library:

  http://work.caltech.edu/library

- To discuss the final, you are encouraged to take part in the discussion forum. Please make sure you don't discuss specific answers, or specific excluded answers, before the final exam is due.

Nonlinear transforms

1. The polynomial transform of order $Q = 10$ applied to $\mathcal{X}$ of dimension $d = 2$ results in a $\mathcal{Z}$ space of what dimensionality (not counting the constant coordinate $x_0 = 1$ or $z_0 = 1$)?

[a] 12

[b] 20

[c] 35

[d] 100

1+2+3+4+5+6+7+8+9+10+11 = 66

[e] None of the above

Bias and Variance

2. Recall that the average hypothesis $\bar{g}$ was based on training the same model $\mathcal{H}$ on different data sets $\mathcal{D}$ to get $g^{(\mathcal{D})} \in \mathcal{H}$, and taking the expected value of $g^{(\mathcal{D})}$ w.r.t. $\mathcal{D}$ to get $\bar{g}$. Which of the following models $\mathcal{H}$ could result in $\bar{g} \notin \mathcal{H}$?

[a] A singleton $\mathcal{H}$ ($\mathcal{H}$ has one hypothesis)

[b] $\mathcal{H}$ is the set of all constant, real-valued hypotheses

[c] $\mathcal{H}$ is the linear regression model

[d] $\mathcal{H}$ is the logistic regression model

[e] None of the above

Overfitting

3. Which of the following statements is false

[a] If there is overfitting, there must be two or more hypotheses that have different values of E_{in}

[b] If there is overfitting, there must be two or more hypotheses that have different values of E_{out}

[c] If there is overfitting, there must be two or more hypotheses that have different values of $(E_{\text{out}} - E_{\text{in}})$

[d] We can always determine if there is overfitting by comparing the values of $(E_{\text{out}} - E_{\text{in}})$

[e] We cannot determine overfitting based on one hypothesis only

4. Which of the following statements is true

[a] Deterministic noise cannot occur with stochastic noise

[b] Deterministic noise does not depend on the learning model

[c] Deterministic noise does not depend on the target function

[d] Stochastic noise does not depend on the learning model

~~[e]~~ Stochastic noise does not depend on the target distribution

Regularization

5. The regularized weight $\mathbf{w}_{\text{reg}}$ is a solution to:

$$\text{minimize} \quad \frac{1}{N}\sum_{n=1}^{N}(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n - y_n)^2 \quad \text{subject to} \quad \mathbf{w}^{\mathrm{T}}\Gamma^{\mathrm{T}}\Gamma\mathbf{w} \leq C,$$

where Γ is a matrix. If $\mathbf{w}_{\text{lin}}^{\mathrm{T}}\Gamma^{\mathrm{T}}\Gamma\mathbf{w}_{\text{lin}} \leq C$, where $\mathbf{w}_{\text{lin}}$ is the linear regression solution, then what is $\mathbf{w}_{\text{reg}}$?

[a] $\mathbf{w}_{\text{reg}} = \mathbf{w}_{\text{lin}}$

[b] $\mathbf{w}_{\text{reg}} = \Gamma\mathbf{w}_{\text{lin}}$

[c] $\mathbf{w}_{\text{reg}} = \Gamma^{\mathrm{T}}\Gamma\mathbf{w}_{\text{lin}}$

[d] $\mathbf{w}_{\text{reg}} = C\Gamma\mathbf{w}_{\text{lin}}$

[e] $\mathbf{w}_{\text{reg}} = C\mathbf{w}_{\text{lin}}$

6. Soft-order constraints that regularize polynomial models can be

[a] written as hard-order constraints

[b] translated into augmented error

[c] determined from the value of the VC dimension

[d] used to decrease both E_{in} and E_{out}

[e] None of the above is true

Regularized Linear Regression

We are going to experiment with linear regression for classification on a real world dataset. Download the processed US Postal Service Zip Code dataset with extracted features of symmetry and intensity for training and testing:

`http://www.amlbook.com/data/zip/features.train` http://www.amlbook.com/data/zip/features.test

(the format of each row is: **digit symmetry intensity**). We will train two types of binary classifiers; one-versus-one (one digit is class $+1$ and another digit is class -1, with the rest of the digits disregarded), and one-versus-all (one digit is class $+1$ and the rest of the digits are class -1). When evaluating E_{in} and E_{out}, use binary classification error. Implement the regularized least-squares linear regression for classification that minimizes

$$\frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{w}^{\mathrm{T}}\mathbf{z}_n - y_n\right)^2 \; + \; \frac{\lambda}{N}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

where $\mathbf{w}$ includes w_0.

7. Set $\lambda = 1$ and do not apply a feature transform (i.e., use $\mathbf{z} = \mathbf{x} = (1, x_1, x_2)$). Which among the following classifiers has the lowest E_{in}?

[a] 5 versus all

[b] 6 versus all

[c] 7 versus all

[d] 8 versus all

[e] 9 versus all

```
5  versus all with lambda =  1  the E_in is: 0.0762584007681 and the E_out is:  0.079720976582
6  versus all with lambda =  1  the E_in is: 0.0910711836511 and the E_out is:  0.0847035376183
7  versus all with lambda =  1  the E_in is: 0.0884652311068 and the E_out is:  0.0732436472347
8  versus all with lambda =  1  the E_in is: 0.0743382252092 and the E_out is:  0.0827105132038
9  versus all with lambda =  1  the E_in is: 0.0883280757098 and the E_out is:  0.0881913303438
```

8. Now, apply a feature transform $\mathbf{z} = (1, x_1, x_2, x_1x_2, x_1^2, x_2^2)$, and set $\lambda = 1$. Which among the following classifiers has the lowest E_{out}?

[a] 0 versus all

[b] 1 versus all

[c] 2 versus all

[d] 3 versus all

[e] 4 versus all

```
0  versus all transformed with lambda =  1  the E_in is: 0.10231792621 and the E_out is:  0.106626806178
1  versus all transformed with lambda =  1  the E_in is: 0.0123439857358 and the E_out is:  0.02192326856
2  versus all transformed with lambda =  1  the E_in is: 0.100260595254 and the E_out is:  0.0986547085202
3  versus all transformed with lambda =  1  the E_in is: 0.0902482512687 and the E_out is:  0.0827105132038
4  versus all transformed with lambda =  1  the E_in is: 0.0894253188863 and the E_out is:  0.0996512207275
```

9. If we compare using the transform versus not using it, and apply that to '0 versus all' through '9 versus all', which of the following statements is correct for $\lambda = 1$?

~~**[a]**~~ Overfitting always occurs when we use the transform

~~**[b]**~~ The transform always improves the out-of-sample performance by at least 5% (E_{out} with transform $\leq 0.95 E_{\text{out}}$ without transform)

~~**[c]**~~ The transform does not make any difference in the out-of-sample performance

```
0  versus all with lambda =  1  the E_in is: 0.109312851461 and the E_out is:  0.11509715994
0  versus all transformed with lambda =  1  the E_in is: 0.10231792621 and the E_out is:  0.106626806178
9  versus all with lambda =  1  the E_in is: 0.0883280757098 and the E_out is:  0.0881913303438
9  versus all transformed with lambda =  1  the E_in is: 0.0883280757098 and the E_out is:  0.0881913303438
5  versus all with lambda =  1  the E_in is: 0.0762584007681 and the E_out is:  0.079720976582
5  versus all transformed with lambda =  1  the E_in is: 0.0762584007681 and the E_out is:  0.0792227204783
```

~~[d]~~ The transform always worsens the out-of-sample performance by at least 5%

[e] The transform improves the out-of-sample performance of '5 versus all,' but by less than 5%

10. Train the '1 versus 5' classifier with $\mathbf{z} = (1, x_1, x_2, x_1x_2, x_1^2, x_2^2)$ with $\lambda = 0.01$ and $\lambda = 1$. Which of the following statements is correct?

[a] Overfitting occurs (from $\lambda = 1$ to $\lambda = 0.01$)

[b] The two classifiers have the same E_{in}

[c] The two classifiers have the same E_{out}

[d] When λ goes up, both E_{in} and E_{out} go up

[e] When λ goes up, both E_{in} and E_{out} go down

```
1  versus  5  transformed with lambda =  1  the E_in is: 0.00512491992313 and the E_out is:  0.0259433962264
1  versus  5  transformed with lambda =  0.01  the E_in is: 0.00448430493274 and the E_out is:  0.0283018867925
```

Support Vector Machines

11. Consider the following training set generated from a target function $f : \mathcal{X} \to \{-1, +1\}$ where $\mathcal{X} = \mathbb{R}^2$

$$\begin{array}{ccc} \mathbf{x}_1 = (1,0), y_1 = -1 & \mathbf{x}_2 = (0,1), y_2 = -1 & \mathbf{x}_3 = (0,-1), y_3 = -1 \\ \mathbf{x}_4 = (-1,0), y_4 = +1 & \mathbf{x}_5 = (0,2), y_5 = +1 & \mathbf{x}_6 = (0,-2), y_6 = +1 \\ & \mathbf{x}_7 = (-2,0), y_7 = +1 & \end{array}$$

Transform this training set into another two-dimensional space $\mathcal{Z}$

$$z_1 = x_2^2 - 2x_1 - 1 \qquad z_2 = x_1^2 - 2x_2 + 1$$

Using geometry (not quadratic programming), what values of $\mathbf{w}$ (without w_0) and b specify the separating plane $\mathbf{w}^{\mathrm{T}}\mathbf{z} + b = 0$ that maximizes the margin in the $\mathcal{Z}$ space? The values of w_1, w_2, b are:

[a] $-1,\ 1,\ -0.5$

[b] $1,\ -1,\ -0.5$

[c] $1,\ 0,\ -0.5$

[d] $0,\ 1,\ -0.5$

[e] None of the above would work

12. Consider the same training set of the previous problem, but instead of explicitly transforming the input space $\mathcal{X}$, apply the SVM algorithm with the kernel

$$K(\mathbf{x}, \mathbf{x}') = (1 + \mathbf{x}^{\mathrm{T}}\mathbf{x}')^2$$

(which corresponds to a second-order polynomial transformation). Set up the expression for $\mathcal{L}(\alpha_1 ... \alpha_7)$ and solve for the optimal $\alpha_1, ..., \alpha_7$ (numerically, using a quadratic programming package). The number of support vectors you get is in what range?

[**a**] 0-1

[**b**] 2-3

[✓ [**c**] 4-5]

[**d**] 6-7

[**e**] >7

Radial Basis Functions

We experiment with the RBF model, both in regular form (Lloyd + pseudo-inverse) with K centers:

$$\text{sign}\left(\sum_{k=1}^{K} w_k \ \exp\left(-\gamma \ ||\mathbf{x} - \mu_k||^2\right) + \ b \right)$$

(notice that there is a bias term), and in kernel form (using the RBF kernel in hard-margin SVM):

$$\text{sign}\left(\sum_{\alpha_n > 0} \alpha_n y_n \ \exp\left(-\gamma \ ||\mathbf{x} - \mathbf{x}_n||^2\right) + \ b \right).$$

The input space is $\mathcal{X} = [-1, 1] \times [-1, 1]$ with uniform probability distribution, and the target is

$$f(\mathbf{x}) = \text{sign}(x_2 - x_1 + 0.25 \sin(\pi x_1))$$

which is slightly nonlinear in the $\mathcal{X}$ space. In each run, generate 100 training points at random using this target, and apply both forms of RBF to these training points. Here are some guidelines:

- Repeat the experiment for as many runs as needed to get the answer to be stable (statistically away from flipping to the closest competing answer).
- In case a data set is not separable in the '$\mathcal{Z}$ space' by the RBF kernel using hard-margin SVM, discard the run but keep track of how often this happens, if ever.
- When you use Lloyd's algorithm, initialize the centers to random points in $\mathcal{X}$ and iterate until there is no change from iteration to iteration. If a cluster becomes empty, discard the run and repeat.

13. For $\gamma = 1.5$, how often do you get a data set that is not separable by the RBF kernel (using hard-margin SVM). *Hint: Run the hard-margin SVM, then check that the solution has $E_{\text{in}} = 0$.*

[a] $\leq 5\%$ of the time

[b] $> 5\%$ but $\leq 10\%$ of the time

```
0.034, 0.051, 0.051, 0.049, 0.043, 0.037, 0.046, 0.04, 0.062, 0.046]
```

[c] $> 10\%$ but $\leq 20\%$ of the time

[d] $> 20\%$ but $\leq 40\%$ of the time

[e] $> 40\%$ of the time

14. If we use $K = 9$ for regular RBF and take $\gamma = 1.5$, how often does the kernel form beat the regular form (excluding runs mentioned in Problem 13 and runs with empty clusters, if any) in terms of E_{out}?

```
0.153
0.161
0.16400000000000001
0.16400000000000001
```

~~[a]~~ $\leq 15\%$ of the time

~~[b]~~ $> 15\%$ but $\leq 30\%$ of the time

[c] $> 30\%$ but $\leq 50\%$ of the time

Second chance:

```
0.104
0.098000000000000004
```

[d] $> 50\%$ but $\leq 75\%$ of the time

[e] $> 75\%$ of the time

15. If we use $K = 12$ for regular RBF and take $\gamma = 1.5$, how often does the kernel form beat the regular form (excluding runs mentioned in Problem 13 and runs with empty clusters, if any) in terms of E_{out}?

```
0.184
```

~~[a]~~ $\leq 10\%$ of the time

Second Chance:

```
0.087999999999999995
0.080000000000000002
```

~~[b]~~ $> 10\%$ but $\leq 30\%$ of the time

[c] $> 30\%$ but $\leq 60\%$ of the time

[d] $> 60\%$ but $\leq 90\%$ of the time

[e] $> 90\%$ of the time

16. Now we focus on regular RBF only, with $\gamma = 1.5$. If we go from $K = 9$ clusters to $K = 12$ clusters (only 9 and 12), which of the following 5 cases happens most often in your runs (excluding runs with empty clusters, if any)?

[a] E_{in} goes down but E_{out} goes up

[b] E_{in} goes up but E_{out} goes down

```
(E_in_RBF_9 < E_in_RBF_12, E_out_RBF_9 < E_out_RBF_12)
array([ 0.34,  0.34])
```

[c] Both E_{in} and E_{out} go up

```
array([ 0.642,  0.64 ])
```

[d] Both E_{in} and E_{out} go down

[e] There is no change

17. For regular RBF with $K = 9$, if we go from $\gamma = 1.5$ to $\gamma = 2$ (only 1.5 and 2), which of the following 5 cases happens most often in your runs (excluding runs with empty clusters, if any)?

[a] E_{in} goes down but E_{out} goes up

~~[b]~~ E_{in} goes up but E_{out} goes down

[c] Both E_{in} and E_{out} go up

[d] Both E_{in} and E_{out} go down

[e] There is no change

(E_in_RBF_1 < E_in_RBF_2, E_out_RBF_1 < E_out_RBF_2)
array([0.972, 0.388])

second chance:
array([0.268, 0.292])

18. What is the percentage of time that regular RBF achieves $E_{\text{in}} = 0$ with $K = 9$ and $\gamma = 1.5$ (excluding runs with empty clusters, if any)?

[a] $\leq 10\%$ of the time

0.002

[b] $> 10\%$ but $\leq 20\%$ of the time

[c] $> 20\%$ but $\leq 30\%$ of the time

[d] $> 30\%$ but $\leq 50\%$ of the time

[e] $> 50\%$ of the time

Bayesian Priors

19. Let $f \in [0, 1]$ be the unknown probability of getting a heart attack for people in a certain population. Notice that f is just a constant, not a function, for simplicity. We want to model f using a hypothesis $h \in [0, 1]$. Before we see any data, we assume that $P(h = f)$ is uniform over $h \in [0, 1]$ (the prior). We pick one person from the population, and it turns out that he or she had a heart attack. Which of the following is true about the posterior probability that $h = f$ given this sample point?

[a] The posterior is uniform over $[0, 1]$

[b] The posterior increases linearly over $[0, 1]$

[c] The posterior increases nonlinearly over $[0, 1]$

~~[d]~~ The posterior is a delta function at 1 (implying f has to be 1)

~~[e]~~ The posterior cannot be evaluated based on the given information

Aggregation

20. Given two learned hypotheses g_1 and g_2, we construct the aggregate hypothesis g given by $g(\mathbf{x}) = \frac{1}{2}\left(g_1(\mathbf{x}) + g_2(\mathbf{x})\right)$ for all $\mathbf{x} \in \mathcal{X}$. If we use mean-squared error, which of the following statements is true?

~~**[a]**~~ $E_{\text{out}}(g)$ cannot be worse than $E_{\text{out}}(g_1)$

[b] $E_{\text{out}}(g)$ cannot be worse than the smaller of $E_{\text{out}}(g_1)$ and $E_{\text{out}}(g_2)$

[c] $E_{\text{out}}(g)$ cannot be worse than the average of $E_{\text{out}}(g_1)$ and $E_{\text{out}}(g_2)$

[d] $E_{\text{out}}(g)$ has to be between $E_{\text{out}}(g_1)$ and $E_{\text{out}}(g_2)$ (including the end values of that interval)

~~**[e]**~~ None of the above