Machine Learning 1 — mock exam

Preliminaries

- The exam is open book. You may use all the material you want, while obeying the following rules:
  - you are not allowed to consult or communicate with other people, be it in the room or anywhere outside, except for with the examiners;
  - you must always place the screens of your computers and other used digital devices so that the examiners can see what you are doing;
  - failure to comply with these simple rules may lead to 0 points.

  In short, we will be as fair as we can, but expect the same from you in return.
- Where requested to draw in figures, do so on the exam printouts.
- This mock exam is limited to 100 minutes. The questions are laid out for 180 minutes. If you can solve 50% of the questions, you are in a good shape.
- You can earn up to 95 points for this exam. The total points in this exam is 95 points. You will not need more than 95 points total for a perfect exam grade. The points you may have earned in your midterm will be added to the points you may earn here. Your final grade is computed from the sum of the points of the two exams.
- This exam consists of 13 pages, 14 sections, 39 problems.
- This is a mock exam only. You will not receive any credit by doing this exam. Do not hand in anything.

1 THE BEGINNING

Problem 1 [1 point] Write your immat but not your name on every page that you hand in. You will also write on these exam sheets, so write them on these, too, on the line at the bottom of each page directly after the text "immat:".

2 Linear Algebra

Consider the following objective function

$$J(\boldsymbol{w}) = \|\boldsymbol{M} - \boldsymbol{w}\boldsymbol{w}^T\|^2$$

$\boldsymbol{M} \in \mathbb{R}^{n \times n}$ is a symmetric positive definite matrix, $\boldsymbol{w} \in \mathbb{R}^n$ is an n-dimensional vector, the squared norm $\|\boldsymbol{A}\|^2$ of a matrix $\boldsymbol{A}$ is defined as $\|\boldsymbol{A}\|^2 = \sum_{ij} A_{ij}^2$.

Problem 2 [2 points] Show that $\sum_{ij} A_{ij}B_{ij} = tr(\boldsymbol{A}^T\boldsymbol{B}) = tr(\boldsymbol{A}\boldsymbol{B}^T)$. $\boldsymbol{A}$ and $\boldsymbol{B}$ are matrices with suitable dimensions.

The trace operator takes the sum of the diagonal elements of a matrix. The j-th diagonal element of $\boldsymbol{A}^T\boldsymbol{B}$ is $\sum_i A_{ij}B_{ij}$

Problem 3 [2 points] Show that $J(\boldsymbol{w}) \equiv \|\boldsymbol{w}\|^4 - 2\boldsymbol{w}^T\boldsymbol{M}\boldsymbol{w} + tr(\boldsymbol{M}\boldsymbol{M})$. Hint: Use the relation just proven.

Using $\|\boldsymbol{A}\|^2 = \sum_{ij} A_{ij}^2$ we can write $J(\boldsymbol{w}) = \sum_{ij}(\boldsymbol{M} - \boldsymbol{w}\boldsymbol{w})_{ij}^2$. This is the (ij)–th element squared. $(\boldsymbol{M} - \boldsymbol{w}\boldsymbol{w})_{ij} = \boldsymbol{M}_{ij} - \boldsymbol{w}_i\boldsymbol{w}_j$, i.e. note to take the element operation before the square operation. Squaring this out gives: $(\boldsymbol{M} - \boldsymbol{w}\boldsymbol{w})_{ij}^2 = \boldsymbol{M}_{ij}^2 - 2\boldsymbol{M}_{ij}\boldsymbol{w}_i\boldsymbol{w}_j + \boldsymbol{w}_i^2\boldsymbol{w}_j^2$. What is missing for $J(\boldsymbol{w})$ is the sum over all (ij) pairs, which can be done for every term independently: $\sum_{ij} \boldsymbol{M}_{ij}^2 = tr(\boldsymbol{M}\boldsymbol{M})$, $\sum_{ij} \boldsymbol{M}_{ij}\boldsymbol{w}_i\boldsymbol{w}_j = \boldsymbol{w}^T\boldsymbol{M}\boldsymbol{w}$ and $\sum_{ij} \boldsymbol{w}_i^2\boldsymbol{w}_j^2 = (\sum_i \boldsymbol{w}_i^2)(\sum_j \boldsymbol{w}_j^2) = \|\boldsymbol{w}\|^2\|\boldsymbol{w}\|^2$.

Problem 4 [3 points] What is the gradient of $J(\boldsymbol{w})$ with respect to $\boldsymbol{w}$? Hint: Use the relation just proven.

The derivative of $\|\boldsymbol{w}\|$ with respect to $\boldsymbol{w}$ is $2\boldsymbol{w}$, you can check this for example by elementwise differentiation. Then $\nabla_{\boldsymbol{w}} J(\boldsymbol{w}) = 8\|\boldsymbol{w}\|^3\boldsymbol{w} - 4\boldsymbol{M}\boldsymbol{w}$.

Problem 5 [4 points] Show that if this gradient is zero for some vector $\boldsymbol{v}$, then $\boldsymbol{v}$ is an eigenvector of $\boldsymbol{M}$.

If $\nabla_{\boldsymbol{w}} J(\boldsymbol{v}) \equiv 0$ then $\boldsymbol{M}\boldsymbol{v} = 2\|\boldsymbol{v}\|^3\boldsymbol{v}$, which corresponds to the defintion of an eigenvector.

3 Probability Theory

Problem 6 [3 points] Your good friend Mark Z. has a safe with three coins in it. One coin is completely golden, one coin is completely made out of platinum and one coin has a golden side and a platinum side. The coins don't have any labels (like a *Head* or *Tail* side). In one of your recent meetings with him he drew a coin randomly from the safe, tossed it and showed you that it has a golden side. Assuming that each coin is fair (i.e., the chance that a coin ends up on either side is 1/2), what is the probability that this is the golden coin (i.e., the other side is golden, too)? Show your work!

We need to compute $p(gg|g)$, i.e. the probablity that both sides are golden, given that one side is golden. Bayes tells us that this is $p(gg|g) = p(g|gg)p(gg)/p(g)$. Clearly, $p(g|gg) = 1, p(gg) = 1/3$, the one coin was drawn randomly! $p(g)$ is the probability that a side of a coin is golden, so apart from the options just presented there is also $p(g|gp) * p(gp) = 1/2 * 1/3$. The coin with two platinum sides never will show a golden side. Thus we have $p(g) = 1/3 + 1/6$ and therefore $p(gg|g) = 2/3$.

4 Inequalities are tricky, or: From Jensen with love

Problem 7 [5 points] Too often did you hear that knowing the uncertainty of some observation is a valuable information. To get rid of that once and for all, you came up with an extremely cool estimator $\hat{\sigma}^2$ for the variance (denoted by σ^2) of any stochastic process. Even better, this estimator is provably unbiased!

Mr. Sam One comes along and decides to use your estimator to compute the *precision* (denoted by τ) of a stochastic process he currently works on: He simply follows the definition of precision ($\tau = 1/\sigma^2$) and sets $\hat{\tau} := 1/\hat{\sigma}^2$. After lots of fiddling it turns out that $\hat{\tau}$ seems to be incorrect. Sam thinks that your variance estimator consistently *underestimates* the variance (i.e., your estimated variance is too small). Explain this observation. Provide all necessary intermediate computation steps!

Trying to decipher what is written here: There is a procedure (an estimator) that computes for some observations a number (denoted $\hat{\sigma}^2$). If we do this for several batches of observations from the same process (a thing that produces observations), then in the mean (simplifying here!) this number matches the true equivalent number (representing an abstract property) of our process (i.e. we have an unbiased estimator). Written in formula: $E[\hat{\sigma}^2] = \sigma^2$. Now, somebody is interested in another property of our process, which is actually a function of the original property, in our case the inverse of the variance. To compute an estimate of this property out of observations, this somebody might be interested in reusing already done work, after all the computation seems straight forward. The question now tells us that in this specific case the unbiasedness does not transfer over, that is, computing a new property out of an old property (an indirect computation, instead of coming up with a new formula for this property that uses the observations directly) seems incorrect. The question even more tells us, that something goes consistently wrong.

The title already hints at using Jensen's inequality. Following this hint, where is a convex function involved here? It is $1/x$.

$$E[\hat{\tau}] = E[\frac{1}{\hat{\sigma}^2}] \geq \frac{1}{E[\hat{\sigma}^2]} = \frac{1}{\sigma^2} = \tau$$

The $\geq$ comes from Jensen. So this tells us that the estimated $\hat{\tau}$ will always be greater (or equal) to the real τ. Somebody who is looking at finding mistakes with other people will then conclude that the original estimator must have been consistently wrong: $\hat{\sigma}^2 \leq \sigma^2$.

5 Mark, you, and the coin

Once again, you're sitting in Mark's office. This time however, it is not Mark who sits in the nice leather chair, no, it is his speaking parrot Zucky. Mark himself is having a relaxing bath in his personal spa right next to the office. You can't see him, but because of a funny *bling* noise coming from the spa you know that Mark is engaged in his favorite past time, tossing gold coins. Mark shouts: "Dude, last time we talked, you seemed to have a knack for golden coin tossing, that's why you are here. I wonder if this coin here is biased. Let me flip it a couple of times for you and you tell me what you think!" He starts tossing (according to the *blings*) so you shout back: "Yo, Mark, you know, would be nice if I could *see* every toss... Maybe you can shout what every toss results in?" Mark starts: "Ok, ehrrr, no, sorry, need to call my friend Larry. Let's do it like this then: I toss, and Zucky will tell you the result. Oh, wait, right, Zucky finds it funny to tell sometimes, ehrrr, not the truth, don't know where he picked that habit. Check out the sample run I did with him yesterday, it is the paper lying

right next to you. I'll start tossing when you're done with your math magic, just let me know, Zucky and I are waiting. Yo, Larry, . . . "

You sort your thoughts and start modeling: Denote with f the result of a coin flip ($f = 0$ is heads, $f = 1$ is tails). Model the bias of the coin with θ_1 and use θ_2 for Zucky's *truthness*. Zucky's answer is denoted by z. Furthermore, assume that θ_2 is independent of f and θ_1. Thus, $p(z|f, \theta_2)$ is given as:

	$z = 0$	$z = 1$
$f = 0$	θ_2	$1 - \theta_2$
$f = 1$	$1 - \theta_2$	θ_2

Problem 8 [2 points] Make a *similar* 2×2 table for the joint probability distribution $p(f, z|\boldsymbol{\theta})$ in terms of $\boldsymbol{\theta} = (\theta_1, \theta_2)$. Show your work.

	$z = 0$	$z = 1$
$f = 0$	$\theta_2\theta_1$	$(1 - \theta_2)\theta_1$
$f = 1$	$(1 - \theta_2)(1 - \theta_1)$	$\theta_2(1 - \theta_1)$

Problem 9 [3 points] The sample run $\mathcal{D}$ on the paper looks like this:

f	1	1	0	1	1	0	0
z	1	0	0	0	1	0	1

What are the maximum likelihood estimates for θ_1 and θ_2? Justify your answer. Note that the likelihood function $p(f, z|\theta_1, \theta_2)$ factorises, i.e., $p(f, z|\theta_1, \theta_2) = p(z|f, \theta_2)p(f|\theta_1)$.

The log likelihood of the dataset is $4 \log \theta_2 + 3 \log (1 - \theta_2) + 4 \log (1 - \theta_1) + 3 \log \theta_1$. Thus the MLE of θ_1 is $3/7$ and of θ_2 is $4/7$.

Problem 10 [4 points] Consider a different model for the joint distribution: It has 4 parameters, $\boldsymbol{\theta} = (\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11})$, with $p(f, z|\boldsymbol{\theta}) = \theta_{fz}$. Given the previous dataset $\mathcal{D}$, what is the MLE of $\boldsymbol{\theta}$? Show your work.

Writing down the log likelihood for this model, we get to $2 \log \theta_{00} + 2 \log \theta_{10} + \log \theta_{01} + 2 \log \theta_{11}$. In order to compute the MLE for the 4 parameters, we need to consider that they need to add up to 1 — this is done via a Lagrange multiplier. We get $\theta_{00} = 2/7$, $\theta_{10} = 2/7$, $\theta_{01} = 1/7$ and $\theta_{11} = 2/7$.

6 Linear Regression

Consider the following (scalar, i.e., x, w are scalar) linear regression model

$$z = wx + \epsilon$$

with $\epsilon \sim \mathcal{N}(0, (\sigma x)^2)$ (note the x in the variance!). The value of σ is assumed to be known. We can write the model more compactly as $z \sim \mathcal{N}(wx, \sigma^2 x^2)$.

Problem 11 [2 points] For $w = 1$ how would a plot for sampled pairs (x, z) look like? *Sketch* the plot for $1 \leq x \leq 3$.

Because $z = wx$ and some noise, the basic plot is a diagonal line through the first quadrant. The samples are spread out along this line, and the spread increases quadratically with x (i.e. when going to the right).

Problem 12 [3 points] What is the distribution of z/x if x is given? Show your work!

Because $z \sim \mathcal{N}(wx, \sigma^2 x^2)$ and x is considered constant, we must have $z/x \sim \mathcal{N}(w, \sigma^2)$

Consider the n samples $\{(x_1, z_1), (x_2, z_2), \ldots, (x_n, z_n)\}$. Each x_i is chosen randomly and z_i is subsequently sampled from $z \sim \mathcal{N}(w^* x_i, \sigma^2 x_i^2)$. Here, w^* is the true underlying parameter value—the value of σ^2 is again given and the same as in our model.

Problem 13 [3 points] What is the maximum likelihood estimate of w? You might want to use your answer to problem 12. Show your work.

Because $z/x \sim \mathcal{N}(w^*, \sigma^2)$, w will be the estimate of the mean of a Gaussian. The MLE estimate of the mean of a Gaussian is the average of the collected samples, i.e. $1/n \sum_i z_i/x_i$.

Problem 14 [2 points] What is the variance of your estimate? Show your work.

Applying the rule for the variance of indepentend random variables (z_i/x_i are mutually independent), the variance of the MLE estimate is σ^2/n.

7 Logistic Regression

Consider the two-dimensional data set $\mathcal{D}$ in Figure 1. The data set comprises two different classes. Suppose we want to model the data with the following logistic regression model:

$$p(y = 1|\boldsymbol{x}, \boldsymbol{w}) = \sigma(w_0 + w_1 x_1 + w_2 x_2)$$

with $\boldsymbol{w} = (w_0, w_1, w_2)$ and $\boldsymbol{x} = (x_1, x_2)$. Denote with $\ell(\boldsymbol{w})$ the log likelihood for a given $\boldsymbol{w}$ on the training set $\mathcal{D}$.

Problem 15 [3 points] Suppose we fit the model by the maximum likelihood approach, e.g., minimise

$$J(\boldsymbol{w}) = -\ell(\boldsymbol{w})$$

Sketch a possible decision boundary (corresponding to $\boldsymbol{w}_{MLE}$) for $J(\boldsymbol{w})$ into Figure 1. Is your decision boundary (corresponding to the MLE solution) unique, or are there several ones? Justify your answer! How many classification errors does your method make on the training set?

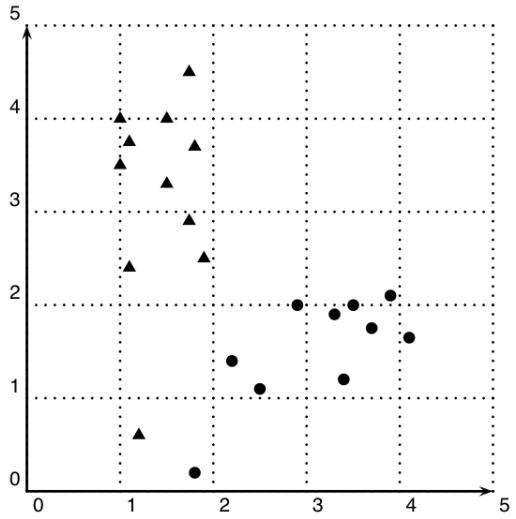


Figure 1: The two classes are labeled with dots and triangles, respectively. **Look carefully!** Use this figure to draw your decision boundaries. x_1 is the horizontal axis, x_2 the vertical axis.

The diagram shows a clear seperation of the two classes. The hyperplane (i.e. line) representing the MLE will take course in this seperating space. It is however not unique — you can draw several lines that make no mistakes. More formally, because you can seperate the two classes, the log likelihood is not bounded and thus there exists no unique optimum (compare to one of the homeworks, where you showed that the weight vector increases in length ad infinitum).

Now, consider that we regularise the bias parameter:

$$J_0(\boldsymbol{w}) = -\ell(\boldsymbol{w}) + \lambda w_0^2$$

Problem 16 [2 points] What is the solution as $\lambda \to \infty$? Again, sketch a possible decision boundary for $J_0(\boldsymbol{w})$. Make sure that the important aspects of the decision boundary are evident.

This means that w_0 will be set to 0. The decision boundary has to go through the origin and thus will make at least one mistake (e.g. $x_1 = x_2$). Again, there are many lines possible.

Now, consider that we regularise *only* the w_1 parameter:

$$J_1(\boldsymbol{w}) = -\ell(\boldsymbol{w}) + \lambda w_1^2$$

Problem 17 [2 points] What is the solution as $\lambda \to \infty$? Again, sketch a possible decision boundary for $J_1(\boldsymbol{w})$ into Figure 1.

The resulting hyperplane is a horizontal line (w_1 will be set to 0), around $x_2 = 2$.

Now, consider that we regularise *only* the w_2 parameter:

$$J_2(\boldsymbol{w}) = -\ell(\boldsymbol{w}) + \lambda w_2^2$$

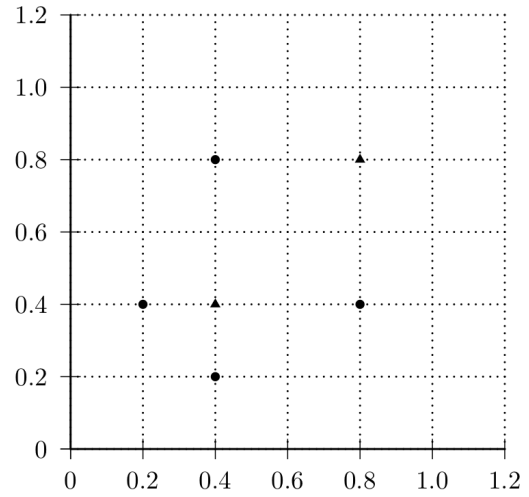
Problem 18 [2 points] What is the solution as $\lambda \to \infty$? Again, sketch a possible decision boundary for $J_2(\boldsymbol{w})$ into Figure 1.

The resulting hyperplane will be a horizontal line, around $x_1 = 2$.

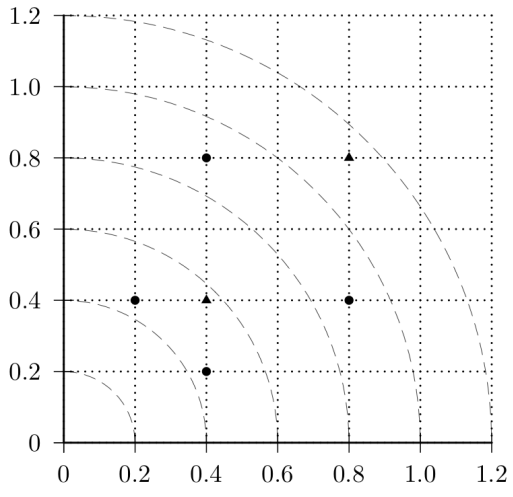
General note for this section: If you are uncertain whether the important aspects of the respective decision boundaries are evident, describe them with a brief sentence.

8 Feature spaces

Sandy and Mandy are chewing over the dataset in Figure 2a. They wonder if it is linearly separable.



(a) Use this figure to draw the final decision boundary.



(b) Use this figure for the new feature space, implicitly defined by K (see text).

Figure 2: The data set of Sandy and Mandy.

Problem 19 [2 points] Is the dataset linearly separable? Justify your answer (you shouldn't need more than one sentence)!

No, because the convex hulls intersect.

Their friend Candy chimes in: "Girls, stop thinking and just throw a kernel on that!" But Sandy and Mandy are not so happy about this idea: "Kernels always mean high dimensions, but 2d is nice, we can draw it on paper." Candy twists here eyes and jots down this function:

$$K(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a}^T \boldsymbol{b}}{||\boldsymbol{a}||\,||\boldsymbol{b}||}$$

Problem 20 [2 points] What is the accompanying feature mapping to this kernel? What is the co-domain of this feature mapping (for $2d$ data like that in Figure 2a)?

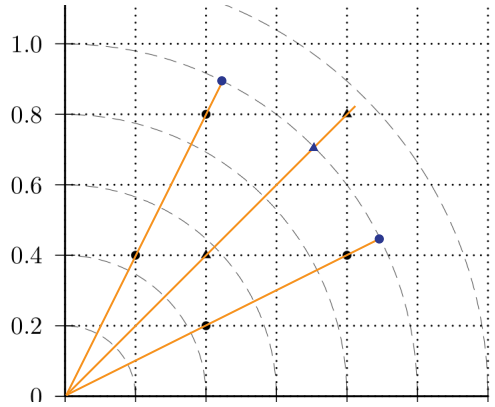
The feature mapping is

$$\Phi(\boldsymbol{x}) = \frac{\boldsymbol{x}}{||\boldsymbol{x}||}\,.$$

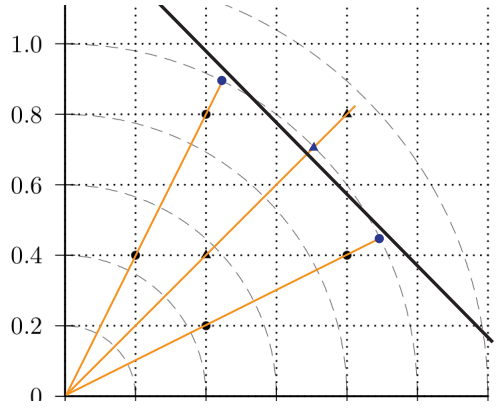
The co-domain of this feature mapping for 2d data is $\mathbb{R}^2$.

Problem 21 [2 points] Map the points to their new representation under this mapping. Use Figure 2b as your feature space.

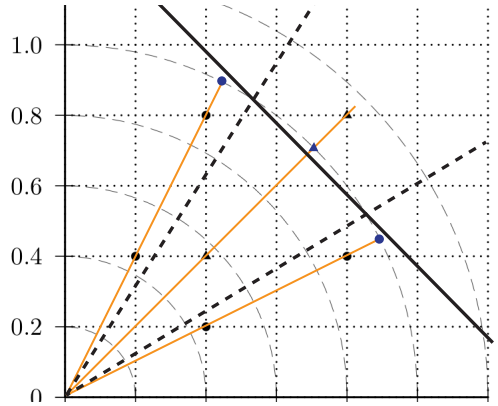
Since the feature map normalizes the $\boldsymbol{x}$ but does not changes the direction of the vector, all samples are projected onto the unit circle.



Problem 22 [2 points] In the same figure (Figure 2b) draw a linear decision boundary that separates the two classes accordingly (it doesn't have to be the one with maximum margin).



Problem 23 [2 points] Furthermore, draw the resulting decision boundary in the original space (use Figure 2a). Be as consistent as possible with your result from problem 22.



9 Mixing binaries

Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ be an n-dimensional *binary* random vector (i.e., $\boldsymbol{x} \in \{0,1\}^n$). The distribution of $\boldsymbol{x}$ is defined as

$$p(\boldsymbol{x}|\boldsymbol{\mu}) = \prod_{i=1}^{n} p(x_i|\mu_i)$$

with $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)$ and

$$p(x_i|\mu_i) = \mu_i^{x_i}(1-\mu_i)^{1-x_i}$$

See the first part of chapter 9.3.3 in Bishop's book (page 445) for the solution to this section.

Problem 24 [2 points] What is the mean and the variance of x_i? Show your work.

Problem 25 [3 points] What is the mean and the covariance matrix of $\boldsymbol{x}$? Show your work.

Now we introduce $\boldsymbol{z} = (z_1, z_2, \ldots, z_n)$, another n-dimensional binary random vector. Its distribution is defined as

$$p(\boldsymbol{z}|\pi_c, \boldsymbol{\mu}_c, c = 1 \ldots C) = \sum_{c=1}^{C} \pi_c p(\boldsymbol{z}|\boldsymbol{\mu}_c)$$

with $\sum_{i=1}^{C} \pi_i = 1$.

Problem 26 [3 points] What is the mean and the covariance matrix of $\boldsymbol{z}$? What is the difference of this covariance matrix to the one of $\boldsymbol{x}$? Show your work.

10 PCA

For an arbitrary dataset, let the empirical covariance matrix $\boldsymbol{\Sigma}$ have eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d > 0$. The *variance of the eigenvalues* σ^2 is defined as $\sigma^2 := \sum_{i=1}^{d}(\lambda_i - \mu_\lambda)^2$, with $\mu_\lambda := \frac{1}{d}\sum_{i=1}^{d}\lambda_i$.

Problem 27 [1 point] What would a low value of σ^2 indicate?

A low σ^2 indicates that all eigenvalues have similar values and thus the principal components take approximately equal part in explaining the variations in the data.

Problem 28 [3 points] Why is σ^2 a good measure of whether or not PCA would be useful for analysing the data?

If σ^2 is high, then some eigenvalues are large while others are small(er) and thus some principal components are more important to explain the variations in the data than others. In this case PCA is useful, because the first few principal components (with the largerst eigenvalues) show what is important in the data and the components with small eigenvalues may be dropped. If σ^2 is low, then all principal components are equally important and PCA just finds another basis for the vector space without much practical significance.

11 Latent Variable Models

Consider the following latent variable model (where $\boldsymbol{z}$ is latent (L-dimensional); y (one-dimensional) and $\boldsymbol{x}$ (d-dimensional) are observed):

$$p(\boldsymbol{z}_i) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_L)$$

$$p(y_i|\boldsymbol{z}_i) = \mathcal{N}(\boldsymbol{w}_y^T \boldsymbol{z}_i + \mu_y, \sigma_y^2)$$

$$p(\boldsymbol{x}_i|\boldsymbol{z}_i) = \mathcal{N}(\boldsymbol{W}_x \boldsymbol{z}_i + \boldsymbol{\mu}_x, \sigma_x^2 \boldsymbol{I}_d)$$

Problem 29 [3 points] Interpret this model verbally (write about 2–5 sentences).

The model looks very similar to a standard pPCA model. Only that there are two seperate visible variables, $\boldsymbol{x}$ and y. Both could be stacked together in order to get the standard pPCA model — note that the equations do not prohibit that. Considering that we often use a multivariate variable ($\boldsymbol{x}$) and a scalar variable (y) as input and ouput in a linear regression model respectively, we could try to apply this view on the model. This then implies (in words) that in such a linear regression model both input and output are connected by a low dimensional ($\boldsymbol{z}$) latent variable. Or, from the viewpoint of dimensionality reduction, when doing dimensionality reduction of $\boldsymbol{x}$, the output y is taken into account.

Problem 30 [4 points] What is $p(y_i|\boldsymbol{x}_i)$? Show your work.

Following the previous argumentation, we stack both $\boldsymbol{x}_i$ and y_i into one vector (denoted, let's say by $\boldsymbol{u}_i$). Thus, from $p(\boldsymbol{u}_i|\boldsymbol{z}_i)$, we can compute the marginal $p(\boldsymbol{u}_i)$, a Gaussian (see the slides on Factor Analysis). We get $p(y_i|\boldsymbol{x}_i)$ by conditioning on $\boldsymbol{x}_i$ in $\boldsymbol{u}_i$ (see the slides on multivariate Gaussians).

$$p(y_i|\boldsymbol{x}_i) = \mathcal{N}(\boldsymbol{w}^T \boldsymbol{x}_i, \sigma_y^2 + \boldsymbol{w}_y^T \boldsymbol{C} \boldsymbol{w}_y)$$

with $\boldsymbol{w} = \boldsymbol{\Psi}^{-1} \boldsymbol{W}_x \boldsymbol{C} \boldsymbol{w}_y$, $\boldsymbol{C}^{-1} = \boldsymbol{I} + \boldsymbol{W}_x^T \boldsymbol{\Psi}^{-1} \boldsymbol{W}_x$ and $\boldsymbol{\Psi} = \sigma_x^2 \boldsymbol{I}_d$.

12 Optimisation

Yaleeza's holiday was boring. Not having access to the internet, she really did not know how to help herself sitting at the beach, listening to the noise of the sea, and observing the apparent rotation of the stars at night. After one week of boredom, she decides to get back to real life. To do this, she points her laptop camera to the sky at night and quickly writes a blob detection algorithm to count the number of visible stars and planets. For a period of 5 consecutive nights, she records the number of stars every 15 minutes, and simultaneously records meteorological data including temperature, humidity, and so on. Using the long periods of darkness in the holiday resort she is staying, she manages to collect 600 data points.

Can she use these data to predict star visibility?

Yaleeza starts off with fitting the data linearly. After normalising all input and output values between -1 and $+1$, she uses a neural network with no hidden units and linear output units ($\phi(x) = x$), and learns the weights using the delta rule. Fitting 600 data points, after a few hundred iterations the training error reduces from 78.3 to 12.2.

Problem 31 [2 points] Write down an equation for the approximation she is learning, i.e., the function $\mathcal{M}(\mathbf{w}, \mathbf{x})$. Use n inputs, one output.

$$\mathcal{M}(\mathbf{w}, \mathbf{x}) \equiv y = \sum_{i=1}^{n} w_i x_i + b$$

Problem 32 [2 points] Could she better have taken another learning method for the approximation you wrote down in the previous problem? Explain your answer.

Yes, for instance SVD. Basically, solving the learning problem means solving a set of linear equations, and—not taking numerical instabilities into account—that involves inverting the matrix W.

Problem 33 [2 points] Did she obtain a good result in learning? What could she have done better?

It is impossible to say if she obtained a good result. The question only gives the approximation error over the training set. Instead, she should have separated the data in a training set and a test set (and a validation set, to be complete). Then she should train on the training set, stop when the error on the test set increases, and test her approximation on the validation set.

After this arduous job, Yaleeza is still uncertain. Is the data well represented? Is this the result she had been looking for? To resolve her doubt, she decides to fit the system with a nonlinear neural network with 5 hidden units. And, indeed, with training a network with $\phi(x) = \tanh(x)$ for the hidden units and linear output units, she ends up with a training error of 9.5 for all 600 samples. She's happy, but not happy enough. Varying the number of hidden units does have an effect, but not as major as she hoped.

Problem 34 [2 points] Write down an equation for the approximation she is learning, i.e., the function $\mathcal{M}(\mathbf{w}, \mathbf{x})$. Use n inputs, m hidden units in one hidden layer, and one output.

For instance,

$$\mathcal{M}(\mathbf{w}, \mathbf{x}) \equiv y = \sum_{i=1}^{m} \tanh \left(\sum_{j=1}^{m} w_{ij} x_i + b_j \right)$$

One can also add biases to the hidden units, but that does not change the representational capabilities of the system.

Problem 35 [3 points] Write out the gradient for weights coming from the input to the hidden and for weights from hidden units to the output unit, with respect to the error function E. Assume that the m hidden units have the above mentioned activation function and there is only one layer of hidden units, n inputs, and one output. Assume a summed squared error function E.

For an output unit j (h is the value of a hidden unit, w'' the weights from hidden to output, w' from input to hidden):

$$\frac{\partial E}{\partial w''_{ji}} = \delta''_j h_i = [z - \mathcal{M}(\mathbf{w}, \mathbf{x})] h_i$$

For a hidden unit j:

$$\frac{\partial E}{\partial w'_{ji}} = \delta'_j x_i = \sum_l \delta''_l w''_{lk} \, x_i$$

13 Generalisation

You are given a dataset containing $n = 1500$ (sample, target) pairs, drawn i.i.d. from an unknown probability density. You are asked to build an optimal model mapping the samples to the targets, and decide to use f-fold cross validation.

Problem 36 [2 points] How many points are used for training and testing when f is 5, 10, 15, 20 in the standard cross-validation setting?

In the required order: $(1200, 300)$, $(1350, 150)$, $(1400, 100)$ and $(1425, 75)$.

Problem 37 [2 points] Now let us assume that the samples were collected in chronological order from a continuous source. In this case, what is the effect of shuffling them before building the cross-validation splits? Is shuffling necessary?

Shuffling is essential, as otherwise the training sets might be biased. For example one training set might contain samples labelled with categories $1, 2, 3, 4$ only, and the testing set might contain samples labelled 5, which would give unpredictable results.

Problem 38 [2 points] Assume the classifier trains in $O(n \log n)$ and the prediction is done in a negligible time. What value of f gives the *fastest* evaluation of the cross-validation error?

Putting ourselves in the worst case and assuming that training will actually take $n \log n$ time units, the evaluation will take $f \cdot |TR| \log |TR|$ where TR is the training set. That means, for $f = 5, 10, 15, 20$, a requirement of about $42500, 97300, 152000, 207000$ time units. When training time dominates over prediction time, the fastest way is that of choosing the smallest possible number of folds.

14 THE END

Problem 39 [1 point] Check if you wrote your immat but never your name on every page that you hand in. You should also have written your immat on these exam sheets, on the line at the bottom of each page directly after the word "immat:".