Probeklausur, Fragen und Antworten

Machine learning (Technische Universität München)

Machine Learning 1 — Mock Exam WS 2017 / 2018

1 Probability Theory

Problem 1 [0 point] You have two coins, C_1 and C_2. Let the outcome of a coin toss be either *heads* ($C_i = 1$) or *tails* ($C_i = 0$) for $i = 1, 2$. C_1 is a fair coin. However, C_2 depends on C_1. If C_1 shows *heads* ($C_1 = 1$), C_2 will show *heads* with probability 0.7. If C_1 shows *tails* ($C_1 = 0$), C_2 will show *heads* with probability 0.5. Now you toss C_1 and C_2 in sequence once. You observe the sum of the two coins $S = C_1 + C_2 = 1$. What is the probability that C_1 shows *tails* and C_2 shows *heads*?

We know:

$$
\begin{aligned}
P(C_1 = H = 1) = P(C_1 = T = 0) &= 0.5 \\
P(C_2 = H = 1 | C_1 = 1) &= 0.7 \\
P(C_2 = H = 1 | C_1 = 0) &= 0.5 \\
S = C_1 + C_2 &= 1
\end{aligned}
$$

We want to know:

$$P(C_1 = 0, C_2 = 1 | S = 1)$$

Therefore, we need Bayes rule [1 point]:

$$
\begin{aligned}
P(C_1, C_2 | S) &\overset{\text{Bayes}}{=} \frac{P(S|C_1, C_2)P(C_1, C_2)}{\text{norm. const.}} \\
&\overset{\text{chain rule}}{=} \frac{P(S|C_1, C_2)P(C_1)P(C_2|C_1)}{\text{norm. const.}} \\
&\overset{\text{expand}}{=} \frac{P(S|C_1, C_2)P(C_1)P(C_2|C_1)}{\sum_{C_1', C_2'} P(S|C_1', C_2')P(C_1')P(C_2'|C_1')}
\end{aligned}
$$

Solve for asked probability [1 point]:

$$\Rightarrow P(C_1 = 0, C_2 = 1 | S = 1) = \frac{\overbrace{P(S = 1 | C_1 = 0, C_2 = 1)}^{=1} \overbrace{P(C_1 = 0)}^{=0.5} \overbrace{P(C_2 = 1 | C_1 = 0)}^{=0.5}}{\sum_{C_1', C_2'} P(S = 1 | C_1', C_2')P(C_1')P(C_2'|C_1')}$$

Expand denominator [1 point]:

$$\begin{aligned}\sum_{C_1',C_2'} P(S=1|C_1',C_2')P(C_1')P(C_2'|C_1') &= \underbrace{P(S=1|C_1'=0,C_2'=0)}_{=0}P(C_1'=0)P(C_2'=0|C_1'=0)\\ &+\underbrace{P(S=1|C_1'=1,C_2'=0)}_{=1}P(C_1'=1)P(C_2'=0|C_1'=1)\\ &+\underbrace{P(S=1|C_1'=0,C_2'=1)}_{=1}P(C_1'=0)P(C_2'=1|C_1'=0)\\ &+\underbrace{P(S=1|C_1'=1,C_2'=1)}_{=0}P(C_1'=1)P(C_2'=1|C_1'=1)\\ &= P(C_1'=1)P(C_2'=0|C_1'=1)+P(C_1'=0)P(C_2'=1|C_1'=0)\\ &= 0.5\times(1-0.7)+0.5\times 0.5\\ &= 0.15+0.25=0.4\end{aligned}$$

Write down final answer [1 point]:

$$\begin{aligned}P(C_1=0,C_2=1|S=1) &= 0.25/0.4\\ &= 0.625 = \frac{5}{8}\end{aligned}$$

2 Parameter Inference / Full Bayesian Approach

For a Naive Bayes classifier we assume the following model:

$$\begin{aligned}p(\mathbf{x},\mathbf{y}|\Theta) &= p(\mathbf{x}|\mathbf{y},\Theta)p(\mathbf{y}|\Theta)\\ &= p(\mathbf{x}|\mathbf{y},\theta,\pi)p(\mathbf{y}|\theta,\pi)\\ &= p(\mathbf{x}|\mathbf{y},\theta)p(\mathbf{y}|\pi)\\ &= \prod_{v=1}^{V} p(x_v|\mathbf{y},\theta)p(\mathbf{y}|\pi)\\ &= \prod_{c=1}^{C}\prod_{v=1}^{V} p(x_v|\theta_{vc})^{y_c}\prod_{c'=1}^{C}\pi_{c'}^{y_{c'}}\end{aligned}$$

where we leave open, which model for the class-conditional densities $p(x_v|\theta_{vc})$ we are using.

Problem 2 [0 point] For this model, write down the posterior distribution for the parameters $p(\Theta|\mathcal{D})$, where $\mathcal{D}=\{\mathbf{x}^{(n)},\mathbf{y}^{(n)}\}_{n=1}^{N}$! It suffices to specify $p(\Theta|\mathcal{D})$ on the $\propto$ level (that is up to constants in Θ) and name the distributions you are introducing as far as the model specification goes.

The likelihood is

$$p(\mathcal{D}|\Theta) = \prod_{n=1}^{N} p(\mathbf{x}^{(n)}, \mathbf{y}^{(n)}|\theta, \pi)$$
$$= \prod_{n=1} [illegible] \prod_{v=1}^{V} p(\mathbf{x}_v^{(n)}|\theta_{vc})^{y_c^{(n)}} \prod_{c'=1}^{C} \pi_{c'}^{y_{c'}^{(n)}}.$$

Assuming suitable co[illegible] $\beta) = p(\theta|\beta)p(\pi|\alpha)$, where for the categorical $p(\mathbf{y}|\pi)$ we should choose a Diric[illegible] get

$$p(\Theta|\mathcal{D}) \propto p(\mathcal{D}|\Theta)p(\Theta|\alpha, \beta)$$

(as always: posterior(Θ) $\propto$ likelihood(Θ) * prior(Θ))

Problem 3 [0 point] Show that for the full Bayesian estimation of the class y for a new data point $\mathbf{x}$ we have

$$p(y_c = 1|\mathbf{x}, \mathcal{D}) \propto \int \prod_{v=1}^{V} p(x_v|\theta_{vc})p(\theta|\mathcal{D})d\theta \int \pi_c p(\pi|\mathcal{D})d\pi$$

$$\begin{aligned}
p(y_c = 1|\mathbf{x}, \mathcal{D}) &= \int p(y_c = 1, \Theta|\mathbf{x}, \mathcal{D})d\Theta \\
&= \int p(y_c = 1|\Theta, \mathbf{x}, \mathcal{D})p(\Theta|\mathbf{x}, \mathcal{D})d\Theta \\
&= \int p(y_c = 1|\Theta, \mathbf{x})p(\Theta|\mathcal{D})d\Theta \\
&\propto \int p(\mathbf{x}|y_c = 1, \Theta)p(y_c = 1|\Theta)p(\Theta|\mathcal{D})d\Theta \\
&\propto \int p(\mathbf{x}|y_c = 1, \theta, \pi)p(y_c = 1|\theta, \pi)p(\theta|\mathcal{D})p(\pi|\mathcal{D})d\theta d\pi \\
&\propto \int p(\mathbf{x}|y_c = 1, \theta)p(\theta|\mathcal{D})d\theta \int p(y_c = 1|\pi)p(\pi|\mathcal{D})d\pi \\
&\propto \int \prod_{v=1}^{V} p(x_v|\theta_{vc})p(\theta|\mathcal{D})d\theta \int \pi_c p(\pi|\mathcal{D})d\pi
\end{aligned}$$

3 Regularized Logistic Regression

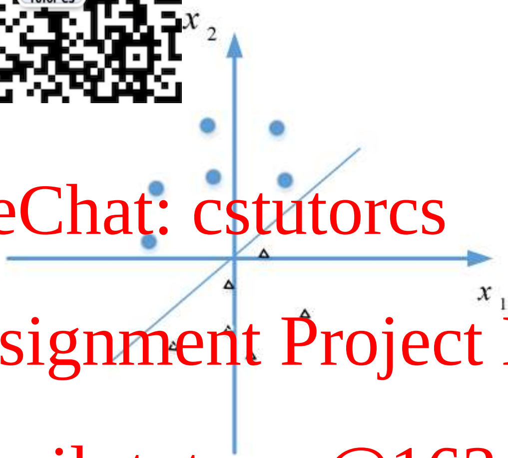
We employ a logistic regression model to classify the data which are plotted in the below figure,

$$\mathbf{p}(y = 1 \mid \mathbf{x}, \mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1 + \exp(\mathbf{w}^T\mathbf{x})}.$$

We fit the data by the maximum likelihood approach, and minimise the negative log-likelihood $-l(\mathbf{w})$, thus the objective function is

$$J(\mathbf{w}) = -l(\mathbf{w}).$$

We get the decision boundary in the figure with zero misclassification error.

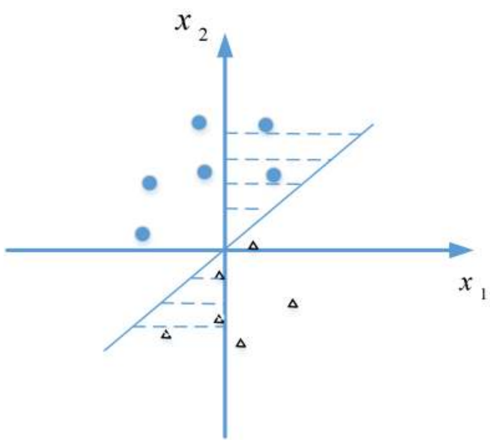


Problem 4 [0 point] Now, we regularise w_2 and minimise

$$J_0(\mathbf{w}) = -l(\mathbf{w}) + \lambda w_2^2.$$

Draw the area that the decision boundary can be in and explain your work.

When we regularise w_2, the decision boundary becomes more vertical. If λ is extremely large, the decision boundary is x_2 axis.



4 Kernels

The following information about kernels *might* be helpful.

Let K_1 and K_2 [illegible] $\mathbb{R}^n$, then the following functions are kernels:

1. $K(\boldsymbol{x}, \boldsymbol{y}) = K$ [illegible]
2. $K(\boldsymbol{x}, \boldsymbol{y}) = \alpha$ [illegible] $\alpha > 0$
3. $K(\boldsymbol{x}, \boldsymbol{y}) = K$ [illegible]
4. $K(\boldsymbol{x}, \boldsymbol{y}) = K$ [illegible] K_3 kernel on $\mathbb{R}^m$ and $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^M$
5. $K(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T B \boldsymbol{y}$ for $B \in \mathbb{R}^{n \times n}$ symmetric and positive semi-definite

Problem 5 [0 point] We have $\mathbf{x} = [x_1\ x_2]^T$. Given the mapping

$$\varphi(x) = [1 \quad x_1^2 \quad \sqrt{2}x_1x_2 \quad x_2^2 \quad \sqrt{2}x_1 \quad \sqrt{2}x_2]^T$$

Determine the kernel $K(\mathbf{x}, \mathbf{y})$. Simplify your answer.

$$K(\mathbf{x}, \mathbf{y}) = \varphi^T(\mathbf{x})\varphi(\mathbf{y}) = (1 + \mathbf{x}^T\mathbf{y})^2$$

Problem 6 [0 point] Let Z be a set of *finite* size. Show that the function

$$K_0(X, Y) = |X \cap Y|$$

is a valid kernel, provided that $X \subseteq Z$ and $Y \subseteq Z$. Remember that Z is finite, i.e. $Z = \{z_1, z_2, \ldots, z_N\}$.

Enumerate all elements of Z, i.e. $Z = \{z_1, z_2, \ldots, z_N\}$. This is possible because Z is of finite cardinality.

Define the feature map $\phi : 2^Z \to \mathbb{R}^N$ by

$$\phi_i(X) = \begin{cases} 1 & \text{if } z_i \in X \\ 0 & \text{if } z_i \notin X \end{cases}.$$

We have

$$K_0(X, Y) = \sum_{i=1}^{N} \underbrace{\phi_i(X)\phi_i(Y)}_{\begin{cases} = 1 & \text{if } z_i \in X \wedge z_i \in Y \\ = 0 & \text{otherwise} \end{cases}} = |X \cap Y|.$$

Problem 7 [0 point] Again, let Z be a set of *finite* size. Show that the function

$$K(X, Y) = 2^{|X \cap Y|}$$

is a valid kernel, provided that $X \subseteq Z$ and $Y \subseteq Z$.

Even if you did not succeed in the previous exercise, you may assume that $K_0(X, Y)$ is a valid kernel.

Set

$$K_1(X, Y) = (\log 2) K_0(X, Y) .$$

This is a kernel (multi... y positive constant).

$\exp(K_1(\boldsymbol{x}, \boldsymbol{y}))$ is a ke...

the Taylor expansion ... nction is

$$\ldots) = 1 + \sum_{n=1}^{\infty} \frac{1}{n!} K_1(\boldsymbol{x}, \boldsymbol{y})^n .$$

The power $K_1(\boldsymbol{x}, \boldsymbol{y})^n$ is a kernel by iterated application of rule 3 ($K_1(\boldsymbol{x}, \boldsymbol{y}) K_2(\boldsymbol{x}, \boldsymbol{y})$ is a kernel). The product $(1/n!) K_1(\boldsymbol{x}, \boldsymbol{y})^n$ is a kernel by rule 2 ($\alpha K_1(\boldsymbol{x}, \boldsymbol{y})$ if a kernel for $\alpha > 0$) because $(1/n!)$ is always positive. The sum $\sum_{n=1}^{\infty} 1/(n!) K_1(\boldsymbol{x}, \boldsymbol{y})^n$ is a kernel by iterated application of rule 1 ($K_1(\boldsymbol{x}, \boldsymbol{y}) + K_2(\boldsymbol{x}, \boldsymbol{y})$ is a kernel). The constant 1 is a kernel by rule 4 ($K_3(\boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{y}))$) with $K_3(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T \boldsymbol{y}$ and $\phi(\boldsymbol{z}) = (1)$. Thus $1 + \sum_{n=1}^{\infty} \frac{1}{n!} K_1(\boldsymbol{x}, \boldsymbol{y})^n$ is a kernel by rule 1.

That $\exp(K_1(\boldsymbol{x}, \boldsymbol{y}))$ is a kernel if $K_1(\boldsymbol{x}, \boldsymbol{y})$ is a kernel was previously shown in the practical session and thus you may thus use this result without re-proving it.

Thus

$$K(X, Y) = \exp((\log 2)|X \cap Y|) = \exp(\log 2)^{|X \cap Y|} = 2^{|X \cap Y|}$$

is a valid kernel.

5 Neural networks

Problem 8 [0 point] Geoffrey has a data set with input $\boldsymbol{x} \in \mathbb{R}^2$ and output $y \in \mathbb{R}^1$. He tests a neural network A with one hidden layer and 9 neurons in that layer (not counting the bias of that layer as a node). He also tests a neural network B with two hidden layers and three neurons for each of these layers (again not counting the biases as nodes). How many free parameters do the two models have? Show your calculation!.

Model A:

- weights from input layer to hidden layer: from each of the two input neurons, we have 9 weights to the hidden layer neurons and we have 9 bias parameters to the hidden neurons. So we have 2 * 9 + 9 parameters here.
- weights from hidden layer to output layer: from each of the 9 hidden neurons we have one weight to the single output neuron and we have one bias parameter to the output neuron. So we have 9 + 1 parameters here.

In total we have 2 * 9 + 9 + 9 + 1 = 37 parameters for model A.

Model B:

- weights from input layer to first hidden layer: from each of the two input neurons, we have 3 weights to the hidden layer neurons and we have 3 bias parameters to the hidden neurons. So we have 2 * 3 + 3 parameters here.
- weights from first hidden layer to second hidden layer: from each of the three first layer

hidden neurons, we have 3 weights to the second hidden layer neurons and we have 3 bias parameters to the second layer hidden neurons. So we have 3 * 3 + 3 parameters here.

- weights from se[illegible] output layer: from each of the 3 second layer hidden neurons we hav[illegible]ngle output neuron and we have one bias parameter to the output neur[illegible]1 parameters here.

In total we have 2 * [illegible]3 + 1 = 25 parameters for model B.

Problem 9 [0 point] Consider a neural network for regression with one output neuron. For that case, the model would be

$$p(y|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(y|y^{NN}(\mathbf{x}, \mathbf{w}), \beta^{-1})$$

where $y^{NN}(\mathbf{x}, \mathbf{w})$ denotes the output of the neural network.

Show that

$$\delta = \frac{\partial E_n}{\partial a} = y^{NN}(\mathbf{x}^{(n)}, \mathbf{w}) - y^{(n)}.$$

Which activation function will you have to use at the output neuron to arrive at this result?

N iid training examples $\{(\mathbf{x}^{(n)}, y^{(n)})\} \to$ likelihood $= \prod_{n=1}^{N} \mathcal{N}(y^{(n)}|y^{NN}(\mathbf{x}^{(n)}, \mathbf{w}), \beta^{-1})$. So (ignoring terms constant in $\mathbf{w}$ and ignoring the β (which is just a multiplicative constant)) we have for the negative log-likelihood

$$NNL(w) = E(w) = \sum_{n=1}^{N} E_n(w) = \sum_{n=1}^{N} \frac{1}{2}(\mathbf{y}^{NN}(\mathbf{x}^{(n)}, \mathbf{w})) - \mathbf{y}^{(n)})^2$$

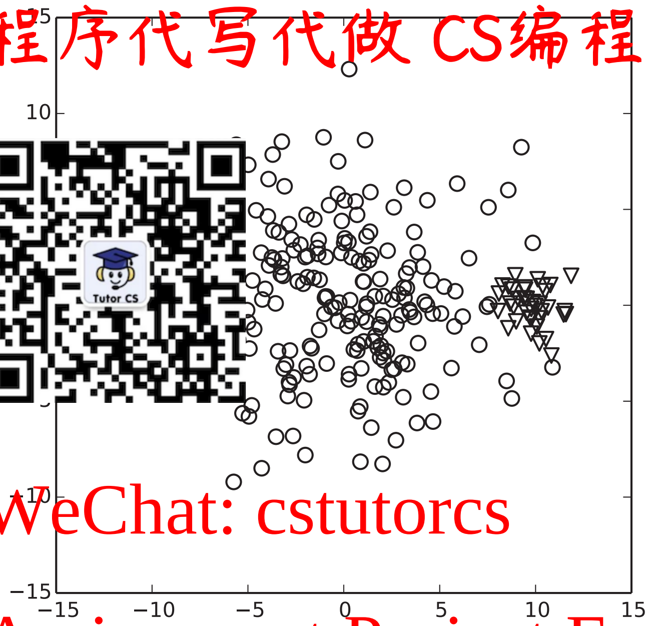
so we get

$$\delta = \frac{\partial E_n}{\partial a} = (y^{NN}(\mathbf{x}^{(n)}, \mathbf{w}) - y^{(n)}) \frac{\partial y^{NN}(\mathbf{x}^{(n)}, \mathbf{w})}{\partial a}.$$

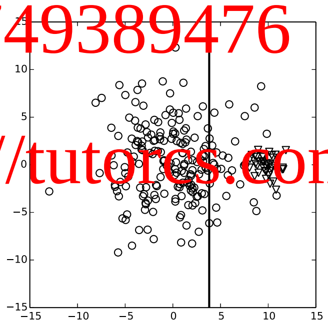
If we use identity as the activation function for the output neuron (that is setting $y^{NN}(\mathbf{x}^{(n)}, \mathbf{w}) = a$), we get the desired result.

6 Clustering

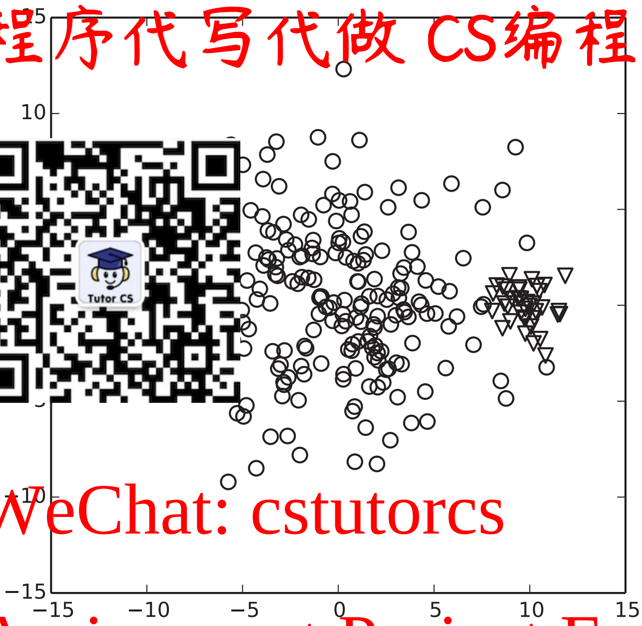
Problem 10 [0 point] Consider the plot below. The data is assumed to have been sampled from two different class-conditional densities and the corresponding class labels are indicated with circles (200 data points) and triangles (40 data points). Now assume that you are given the data of the plot without the class labels. In the plot, draw the resulting decision boundary for cluster assignments for a converged run of k-means (Loyd's algorithm) with two centroids.



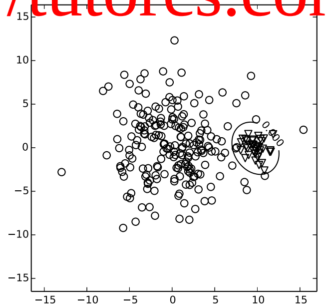
The decision boundary will be a straight line orthogonal to the connecting line between the cluster centroids, separating this line in the middle. Remember that the cluster assignment of k-means does not pay attention to the number of points in each cluster but is only dependent on the cluster centroids. The exact position of the boundary in the plot is not extremely important, but it will not seperate the two classes perfectly.



Problem 11 [0 point] How could we define an analogous hard decision boundary for cluster assignments if instead of k-means (Loyd's algorithm) we would use the EM algorithm with a Gaussian mixture model with two components and individual full covariance matrices as clustering approach? Draw a likely decision boundary qualitatively in the figure!



A reasonable way to define a crisp decision boundary for a pattern $\mathbf{x}^{(n)}$ would be to assign the pattern to cluster 1 if for the responsibilities we have $\gamma_{n1} > \gamma_{n2}$ and else to assign it to cluster 2. After convergence of the EM algorithm, the Gaussian for the first cluster (which would very likely be mainly determined by the data points labeled with squares) will very likely be roughly centered around the points with square labels and will very likely be rather flat and extended over the whole data area. The Gaussian for the second cluster (mainly determined by the data-points labeled with triangles) would very likely have a mean near the centroid of these points with triangle labels and would be rather spherical and more concentrated around the mean. This would imply a decision surface as depcited here:



Problem 12 [0 point] Describe the main steps of the EM algorithm applied to a Gaussian mixture model.

Lecture 11, slide 29

7 Linear Regression

Problem 13 [0 point] For ridge regression, we have the following well known objective function:

$$\ldots w_0\mathbf{1})^T(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1}) + \lambda\mathbf{w}^T\mathbf{w}$$

where $\mathbf{1} = (1, 1, \ldots, 1)^T$ … ust to the lecture slides, we have NOT "absorbed" w_0 into $\mathbf{w}$ by padding each $\mathbf{x}$ w… nponent $= 1$

Assuming $\bar{\mathbf{x}} = 0$, derive the expression for the optimizer for $\mathbf{w}$:

$$\hat{\mathbf{w}}_{ridge} = (\lambda\mathbf{I} + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$J(\mathbf{w}) = \mathbf{y}^T\mathbf{y} + \mathbf{w}^T(\mathbf{X}^T\mathbf{X})\mathbf{w} - 2\mathbf{y}^T(\mathbf{X}\mathbf{w}) + \lambda\mathbf{x}^T\mathbf{x} - 2w_0\mathbf{1}^T\mathbf{y} + 2w_0\mathbf{1}^T\mathbf{X}\mathbf{w} + w_0\mathbf{1}^T\mathbf{1}w_0$$

Due to $\bar{\mathbf{x}} = 0$ we have

$$w_0\mathbf{1}^T\mathbf{X}\mathbf{w} = \bar{\mathbf{x}}^T\mathbf{w} = 0 \tag{1}$$

so we get

$$\frac{\partial}{\partial\mathbf{w}}J(\mathbf{w}) = [2\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{X}^T\mathbf{y}] + 2\lambda\mathbf{w} = 0$$

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

8 Multivariate Gaussian

Problem 14 [0 point] The plot below shows a joint Gaussian distribution $p(x_1, x_2)$. Qualitatively draw the conditionals $p(x_1|x_2 = 0)$ and $p(x_1|x_2 = 2)$ in the given coordinate systems (In the coordinate systems, the vertical axes have an arbitrary scale factor a to avoid having to deal with exact numbers for the vertical axes' values).

Hint: for a general multivariate Gaussian $\mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma})$, where $\mathbf{x} \in \mathbb{R}^D$, the conditional $p(\mathbf{x}_1|\mathbf{x}_2)$ (where we split $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)^T$ into $\mathbf{x}_1 \in \mathbb{R}^M$ and $\mathbf{x}_1 \in \mathbb{R}^{D-M}$) is given by $p(\mathbf{x}_1|\mathbf{x}_2) = \mathcal{N}(\mathbf{x}_1|\mu_{1|2}, \mathbf{\Sigma}_{1|2})$ with $\mu_{1|2} = \mu_1 + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}(\mathrm{x}_2 - \mu_2)$ and $\mathbf{\Sigma}_{1|2} = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$, where $\mu = \begin{pmatrix} \mu_1 \\ \mu_1 \end{pmatrix}$ and $\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}$

For a bivariate Gaussian distribution $p(x_1, x_2) = \mathcal{N}(x_1, x_2|\mu, \Sigma)$ with $\mu = (\mu_1, \mu_2)^T$ and

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$

applying the formula from the hint for a bi-variate Gaussian yields

$$p(x_1|x_2) = \mathcal{N}(x_1|\mu_{1|2}, \sigma_{1|2}) = \mathcal{N}(x_1|\mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(x_2 - \mu_2), \sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2})$$

We see that while $\mu_{1|2}$ depends on the value of x_2, $\sigma_{1|2}$ does not. We do not even need to explicitly derive the above expression to see that: From the hint it is already evident that $\sigma_{1|2}$ is independent of x_2. Since the conditional Gaussian is, of course, a normalized Gaussian, the shape of both conditional Gaussians is thus identical. For the drawing it is sufficient to roughly guess some reasonable value for $\sigma_{1|2}$ from the original plot.

We can further see that for $\mu_1 = \mu_2 = 0$ (which can be seen from the original plot), for the case $x_2 = 0$ we have $\mu_{1|2} = 0$. For general x_2 we do not exactly know σ_{12} and σ_2 from the original plot, but due to the fact the conditional Gaussian is, of course, symmetric around $\mu_{1|2}$, $\mu_{1|2}$ can be inferred graphically as the middle point between the intersections of the horizontal x_2 lines with the iso-curves.

9 Constrained optimization

Find the box with the maximum volume which has surface area no more than $S \in \mathbb{R}^+$.

Problem 15 [0 point] ... an of the problem and the corresponding Lagrange dual function. Hint: set the ... ngth, width and height to be l, w, h respectively.

Problem 16 [0 point] ... blem and give the solution to the original problem. You may assume without pr... gap is zero.

Surface area is $S = 2lw + 2lh + 2hw$ and the volume is lwh

The problem is equivalent to:

Minimize: $f(l, w, h) = -lwh$

Subject to: $h(l, w, h) = lw + lh + hw - \frac{S}{2} \leq 0$

$$L(l, w, h, \alpha) = -lwh + \alpha(lw + lh + hw - \frac{S}{2})$$

Computing the partial derivatives of $L(l, w, h, \alpha)$ with respect to w, l, h gives

$$\frac{\partial L}{\partial l} = -wh + \alpha(w + h) = 0$$

$$\frac{\partial L}{\partial w} = -lh + \alpha(l + h) = 0$$

$$\frac{\partial L}{\partial h} = -wl + \alpha(w + l) = 0$$

Solving this system of equations yields

$$l(\alpha) = w(\alpha) = h(\alpha) = 2\alpha.$$

Inserting this into $L(l, w, h, \alpha)$ yields the Lagrange dual function

$$g(\alpha) = \min_{l,w,h} L(l, w, h, \alpha) = 4\alpha^3 - \alpha\frac{S}{2}.$$

Solving the dual problem:

$$0 = \frac{dg}{d\alpha}$$

subject to dual feasibility $\alpha \geq 0$, yields

$$\alpha = (\frac{S}{24})^{1/2}$$

and thus

$$l = w = h = (\frac{S}{6})^{\frac{1}{2}}$$

$$[illegible]) = -\min(f) = (\frac{S}{6})^{\frac{3}{2}}$$

This was an old exam [illegible]tunately took over for this mock exam without properly checking the solution [illegible]ing the mock exam. There are however a number of problems with the ex[illegible]mendation is to at best forget the whole exercise for your preparation.

First of all regard that the function $f(l, w, h) = -lwh$ is not convex. Furthermore, regard that $l(\alpha) = w(\alpha) = h(\alpha) = 2\alpha$ is not a minimum of L (it is also not a maximum but a saddle point). This can be seen by investigating the eigenvalues of the Hessian matrix This makes inserting $l(\alpha) = w(\alpha) = h(\alpha) = 2\alpha$ into L to yield the Lagrange dual function $g(\alpha)$ pointless in terms of our original formalism. Investigating the the resulting "pseudo" $g(\alpha)$ further reveals that the function is not even concave for $\alpha \geq 0$ (which it should ALWAYS be even if f_0 is not convex. Furthermore, $\alpha = (\frac{S}{24})^{1/2}$ is a MINIMUM of $g(\alpha)$ which is convex for $\alpha \geq 0$.

One way around all this is to argue geometrically that the maximum volume of the box will result if the surface takes its maximum value $2(lw + lh + hw) = S$. Thus we can solve the problem of minimizing $f_0(w, l, h) = -lwh$ by using an EQUALITY constraint $2(lw + lh + hw) = S$ (instead of the inequality constraint $2(lw + lh + hw) \leq S$) and the "same" Lagrangian $L(l, w, h, \alpha) = -lwh + \alpha(lw + lh + hw - \frac{S}{2})$ but this time with the normal procedure of langrage multipliers applied when dealing with equality constraints only.

Thus in addition to

$$\frac{\partial L}{\partial l} = -wh + \alpha(w + h) = 0$$

$$\frac{\partial L}{\partial w} = -lh + \alpha(l + h) = 0$$

$$\frac{\partial L}{\partial h} = -wl + \alpha(w + l) = 0$$

which gives again

$$l(\alpha) = w(\alpha) = h(\alpha) = 2\alpha.$$

we would have to add

$$\frac{\partial L}{\partial \alpha} = 0$$

which (as usual for this procedure) yields back the equality constraint

$$lw + lh + hw - \frac{S}{2} = 0$$

inserting $l(\alpha) = w(\alpha) = h(\alpha) = 2\alpha$ we get

$$12\alpha^2 - \frac{S}{2} = 0$$

and together with $\alpha > 0$ (which stems from the normal procedure of direct constrained optimization with equality constraints) we get the same result as before.

$$\alpha = (\frac{S}{24})^{1/2}$$

So the fact that optimizing the "dual function" $g(\alpha)$ yields the same result may within the boundaries of our usual formalism involving Lagrange dual functions be regarded as a coincidence.

10 Variational Inference

Problem 17 [0 point] Show that evidence lower bound (ELBO), defined as

$$= \mathbb{E}_q\left[\log \frac{p(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right]$$

is a lower bound to the

$$\log p(\mathbf{x}).$$

VI lecture, slides 14-1