Machine Learning 1 WS 2014/15

Gedächtnisprotokoll

6. März 2015

Bearbeitungszeit: 120 Minuten, Punkte: 100.

Aufgabe	Punkte
1	15
2	15
3	25
4	20
5	25

Lösungen sind nicht offiziell, falls du Fehler entdeckst bitte bei der Freitagsrunde melden / Wiki eintragen!

Multi Choice

Aufgabe 1

15 Punkte

Es gibt nur eine richtige Antwort. Falsche Antworten geben 0 Punkte genauso wie keine Antwort.

(a) (3 Punkte) The Bayes error for classification is:

- ☐ the lowest error achievable by a linear classifier.
- ☐ the lowest error achievable by a nonlinear quadratic classifier
- ☐ the lowest error achievable by the best possible classifier
- ☐ the lowest error achievable by a classifier assuming Gaussian-generated distributions.

(b) (3 Punkte) Independent Component Analysis can be achieved by:

- ☐ Finding the directions in the input space that maximize the variance of the projected data.
- ☐ Applying a whitening procedure to the data and running PCA on the whitened data.
- ☐ Finding the directions in the input space that maximize the skewness of the projected data.
- ☐ Finding the directions in the input space that minimize the variance of the projected data.

(c) (3 Punkte) The K-means algorithm:

- ☐ Is a convex algorithm that can be used to cluster the data.
- ☐ Is a nonconvex algorithm that can be used to cluster the data.
- ☐ Is a kernelized version of the means algorithm where the kernel is Gaussian.
- ☐ Is a kernelized version of the means algorithm where the kernel can be arbitrary.

(d) (3 Punkte) A biased estimator is sometimes used to:

- ☐ Reduce the risk of underfitting the data.
- ☐ Reduce the estimation error for high-dimensional data.
- ☐ Make the estimation procedure more sensitive to the observed data.
- ☐ None of the above. We should always favor an unbiased estimator.

(e) (3 Punkte) Which is **False**? The Restricted Boltzmann machine is:

- ☐ A machine learning method that is based on error backpropagation.
- ☐ A machine learning method that can learn initial weights for a neural network.
- ☐ A machine learning method that estimates binary probability distributions for the data.
- ☐ A machine learning method that can learn global and local features in the data.

Lösung:

(a) the lowest error achievable by the best possible classifier

(b) Finding the directions in the input space that maximize the skewness of the projected data.

(c) Is a nonconvex algorithm that can be used to cluster the data.

(d) Reduce the estimation error for high-dimensional data. (James-Stein Estimator)

(e) False: A machine learning method that is based on error backpropagation.

Models and Datasets

Aufgabe 2

15 Punkte

(a) (7 Punkte) Sketch a two-dimensional two-class supervised dataset that Fishers linear discriminant and the linear hard-margin SVM would learn separating bounderies with different directions. Your example must be stereotyping for the two classification techniques.

(b) (7 Punkte) Sketch a two-dimensional unsupervised dataset for which K-means (K=3) may get stuck at a local optimum. In your drawing show (using square markers) the position of the centroids for the local minimum and (with circels) the position of the global minimum. Your example must be stereotyping in order to show the difference between the local and global minimum.

Lösung:

(a) SVM minimiert die margin, Fisher minimiert die variance innerhalb der Klasse. zwei relativ breite wolken führen zu einer schrägen boundary bei fisher, aber einer geraden bei SVM

(b) wenn zwei cluster mit niedriger varianz nah beieinander und einer mit hoher vorhanden sind und der algorithmus zwei startpunkte in dem cluster mit hoher varianz initialisiert wird er davon nicht wegkommen.

Kernels and Feature Maps

Aufgabe 3

25 Punkte

In order for a kernel to be positive semid-definite (PSD) it must satisfy

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k(x_i, x_j) \geq 0$$

for all sequences of data points and coefficients.

(a) (10 Punkte) Show that if k_1 and k_2 are PSD kernels, then $k_3 = \alpha k_1 + \beta k_2$ with $\alpha, \beta \geq 0$ is also a PSD kernel.

(b) (5 Punkte) Give an example showing that the positive semi-definiteness is not guaranteed if $\alpha < 0$ or $\beta < 0$

(c) (10 Punkte) A feature map ϕ associated to a kernel k must satisfy:

$$< \phi(x_i), \phi(x_j) > = \; k(x_i, x_j) \;\; \forall x_i, x_j \in X$$

Find the feature map of $k_3 = \alpha k_1 + \beta k_2$ and show that it fulfills the equation above.

Lösung:

(a) $\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k_3(x_i, x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j(\alpha k_1 + \beta k_2) = \alpha \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k_1 + \beta \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k_2$ each of them is bigger then/equal zero because k_1, k_2 are PSD kernels. The sum of both is still bigger then/equal zero, because the coefficients are non negativ so we add to non negativ numbers.

(b) example: $\alpha = 0, \beta = -1$ Now $\beta \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k_2 < 0$

(c)

$$\phi_3(x) = \begin{pmatrix} \sqrt{\alpha}\phi_1(x) \\ \sqrt{\beta}\phi_2(x) \end{pmatrix}$$

$k_3 = <\phi_3(x_i), \phi_3(x_j)> = \alpha\phi_1(x_i)\phi_1(x_j) + \beta\phi_2(x_i)\phi_2(x_j) = \alpha k_1 + \beta k_2$

Lagrance multipliers

Aufgabe 4

20 Punkte

We consider a discrete probability distribution $p(i), i \in \{1..10\}$. We can represent such probability distribution as a vector p indexed by i subject to the constraints $p_i \geq 0$ and $\sum_{i=1}^{10} p_i = 1$. We would like to find analytically the probability distribution p with maximum entropy.

The Entropy is given by $H(p) = -\sum_{i=1}^{10} p_i \log p_i$ and is a concave function

(a) (5 Punkte) Write down the Lagrangian function associated to this constrained optimization problem.

(b) (10 Punkte) Show using the Lagrangian method that the optimal probability distribution is uniform with $p(i) = 0.1$ for all i.

(c) (5 Punkte) Explain briefly why the same Lagrange method cannot be used to find probability distribution with minimum entropy

Lösung:

(a) $L(p_i, \lambda) = -\sum_{i=1}^{10} p_i \log p_i + \lambda(\sum_{i=1}^{10} p_i - 1)$ The other constrain is omitted.

(b) $\nabla_{p_i} L(p_i, \lambda) = -\log p_i - p_i + \lambda = 0$
for $i \neq j$ follows: $-\log p_i - p_i = -\log p_j - p_j \Rightarrow p_i = p_j$
$\nabla_{\lambda} L(p_i, \lambda) = \sum_{i=1}^{10} p_i - 1 = 0$
therefore $10 \cdot p_i = 1 \Rightarrow p_i = 0.1$

(c) Because Entropy is a concave function, it does not have a global minimum. But we have a constrained problem, so we have a minimum here. the Problem with the Lagrange Method is the derivative at the Boundary of p is not defined.
A result with min Entropy would be: only one p = 1 all other equal zero.
I am not sure if this argumentation is sufficient to get all points.

Quadratic Programming

Aufgabe 5

25 Punkte

We consider the regularized regression task this is solved by optimizing

$$\min_w \sum_{N}^{k=1} (y_k - w^T x_k)^2 \text{ subject to } |w|_\infty \leq 1$$

where w is the solution and $(x, y)_k$ is a dataset of N input-output pairs.

(a) (10 Punkte) Show that the optimization problem can be rewritten as:

$$\min_w w^T X^T X w - 2y^T X w \text{ subject to } w_i \leq 1 \text{ and } w_i \geq -1$$

(b) (15 Punkte) You have at your disposal a quadratic solver v = quadprog(Q,l,A,b) that is solving

$$\min_v v^T Q v + l^T v \text{ subject to } Av \leq b$$

Write down the code that builds the numpy array Q,l,A,b from the data X and y.

Lösung:

(a) $(y_k - w^T x_k)^2 = y_k^2 - 2w^T x_k y_k + (w^T x_k)^2$
Because we minimize with respect to w, we don't care about y_k^2
$\sum_{N}^{k=1} (w^T x_k)^2 - 2w^T x_k y_k = (w^T X^T)^2 - 2y^T Xw = (w^T X^T)^T (w^T X^T) - 2y^T Xw = (Xw)(w^T X^T) - 2y^T Xw = w^T X^T Xw - 2y^T Xw$
I don't know what they expect here for explaination
for the constrains: $|w|_\infty = max\{|w_i|\} \leq 1$ therfore $|w_i| \leq 1 \Leftrightarrow -1 \leq w_i \leq 1$

(b)
```
def regression(X,y):
    Q = X.T.dot(X)
    l = 2*y
    d = Q.shape[1] #number of dimensions
    A = numpy.concatenate((numpy.identiy(d), -1*numpy.identiy(d)), axis=0)
    b = numpy.ones(2*d)
    return quadprog(Q,l,A,b)
```