CS 189 Introduction to Machine Learning
Fall 2017 gs2

1 Backprop: *Adapted from Fall 2015 Final Exam*

Suppose we have a neural network that takes in $P = 2000$ input features, has $H = 500$ hidden units, and $N = 4$ outputs.

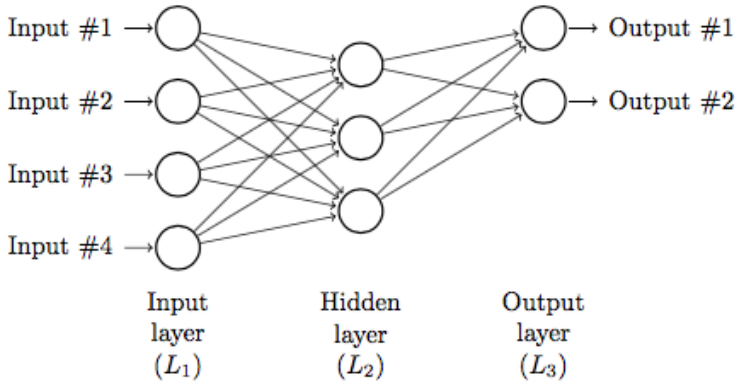


Figure 1: Network without all input and hidden layer nodes shown

$f(x)$ is the activation function in the hidden layer, and $g(x)$ is the activation function in the output layer. You may refer to their respective partial derivatives as $f'(x)$ and $g'(x)$. We will use the mean squared error loss function, which is $L(z) = \frac{1}{2}\|z-y\|^2$ where z is your estimate for label y.

Let's define some notation!
$x \in \mathscr{R}^P$ is a feature vector
$s_j^h = \sum_{i=1}^P V_{i,j} x_i$ are inputs to the hidden layer
$h_j = f(s_j^h)$
$s_k^o = \sum_{j=1}^H W_{j,k} h_j$ are inputs to the output layer
$z_k = g(s_k^o)$
$V \in \mathscr{R}^{H \times P}$ are weights between input layer and hidden layer
$W \in \mathscr{R}^{N \times H}$ are weights between hidden layer and output layer

(a) Suppose you want to include a bias term in both the input layer and hidden layer, how many weights will be trained? How does this change our weight matrices V and W?

(b) Derive the following terms:

$$\frac{\partial L}{\partial W_{j,k}} =$$
$$\frac{\partial L}{\partial V_{i,j}} =$$

(c) As you probably noticed in your homework, vectorizing our backpropagation can speed up computation by taking advantages of matrix operation libraries like numpy. Vectorize the partial derivatives from part (b) and derive the update equation for stochastic gradient descent with learning rate η_1 and η_2 for V and W. (Use $\otimes$ for elementwise multiplication)

(d) The softmax function, or normalized exponential function, is a generalization of the logistic function to handle multiclass classification. The softmax function "squashes" a N-dimensional vector s of arbitrary real values to a K-dimensional vector $\sigma(s)$ of real values in the range [0, 1] that add up to 1. The i-th value corresponds to the probability that the output is class i. The function is given by the following:

$$\sigma(s_j) = \frac{e^{s_j}}{\sum_{k=1}^{N} e^{s_k}} \qquad \text{for } j = 1...N$$

The cross entropy error is commonly paired with softmax activation in the output layer and is defined as:

$$L(z) = \sum_{k=1}^{N} y_k \ln(z_k)$$

Now, let $g(x)$ be the softmax function and let our loss function be the cross entropy loss. Calculate the partial derivative of L with respect to $W_{i,j}$.

2 Kernel PCA

Let $X \in \mathbb{R}^{n \times d}$ be a matrix with rows $x_1^T \ldots x_n^T$, and $\phi : X \longrightarrow X'$ be a feature map with associated kernel $k(x,y) = \langle \phi(x_1), \phi(x_2) \rangle$. We will attempt to perform kernel PCA on X using the feature mapping ϕ. For this problem, assume that the data is already centered.

(a) Let $\Sigma = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T$ be the sample covariance matrix. Recall that performing PCA involves solving for the eigenvectors and eigenvalues of Σ.

Show if v is a solution (an eigenvector of Σ, e.g. $\Sigma v = \lambda v$), then v is in the range of X^T, or $v = X^T \alpha$ for some α.

(b) Show if (α, λ) is a solution to $XX^T \alpha = \lambda \alpha$, then $(v = X^T \alpha, \lambda)$ solves $X^T X v = \lambda v$.

(c) Kernel PCA solves $\phi(\Sigma) v = \lambda v$, where $\phi(\Sigma) = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \phi(x_i)^T$. Explain how to find all λ satisfying this equation without explicitly computing $\phi(x)$.

(d) In regular PCA, the projection coefficient of x onto a principal component v is $\langle x, v \rangle$. In feature space, the coefficient is $\langle \phi(x), v \rangle$. How do we compute this without explicitly computing $\phi(x)$?

(e) To get the correct projection coefficient in regular PCA, we always normalize eigenvectors v so that $\langle x, v \rangle$ gives the correct coefficient. How can we equivalently ensure proper normalization in our kernel PCA?

3 QDA Isocontours

Given the 2×2 matrix

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

(a) Determine the eigenvalues and set of orthonormal eigenvectors of A

(b) Determine the shape of the set given by constraint $x^\top Ax = 1$, where $x \in \mathbb{R}^2$ and $\|x\|_2 = 1$

(c) Recall that the quadratic for an anisotropic distribution of class C is given by the following formula

$$Q_C(x) = -\frac{1}{2}(x-\mu_c)^\top \Sigma_C^{-1}(x-\mu_C) - \frac{1}{2}|\Sigma_C| + \ln \pi_C$$

Assume that we are given two classes: C and D described by the following parameters:

$$\Sigma_C^{-1} = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}, \Sigma_D^{-1} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

$$\mu_C = \mu_D = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\pi_C = \pi_D$$

Also recall that the Bayes decision boundary is the set of points where each class is equally likely. Determine the formula for the Bayes decision boundary.

(d) Assume that the Bayes decision boundary has the following form:

$$(x-\mu)^\top \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} (x-\mu) = 1$$

Draw out the Bayes decision boundary.

(e) Assume now that we are given a new Bayes decision boundary

$$x^\top \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} x = \ln(\frac{4}{3})$$

What would this decision boundary look like?

4 SVM

(a) In a soft margin SVM, if we increase C, which of the following are likely to happen? (Select all that apply)

(a) The margin will grow wider.

(b) All slack variables will go to 0.

(c) The norm $||w||_2$ will grow larger.

(d) There will be more points inside of the margin.

(b) If a hard margin svm tries to minimize $||w||_2^2$ subject to $y_i(w^T x_i + b) \geq c$ for some c, what will the width of the slab (the empty region) be?

(a) $\frac{2}{||w||_2^2}$

(b) $\frac{2c}{||w||_2^2}$

(c) $\frac{c}{||w||_2^2}$

(d) $\frac{2}{c||w||_2^2}$

(c) The shortest distance from a point z to a hyperplane $w^T x = 0$ is

(a) $w^T z$

(b) $\frac{w^T z}{||w||_2}$

(c) $\frac{w^T z}{||w||_2^2}$

(d) $||w||_2||z||_2$

(d) Which of the following is true about SVM's? (Select all that apply)

(a) For soft margin SVM's, a solution exists if and only if the data is linearly separable.

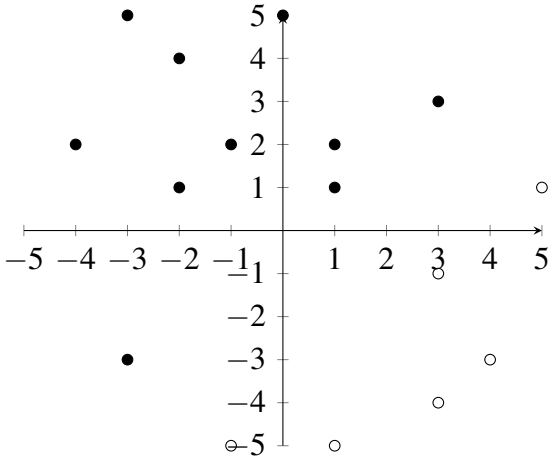
(b) For hard margin SVM's, the support vectors are the only points needed to calculate the decision boundary.

(c) Both hard margin and soft margin SVM's can perfectly separate any training data in some feature space.

(d) Least squares SVM with $l2$-regularization has a closed form solution.

(e) Let's take a look at a set of points and see what a hard margin SVM would determine as the hyperplane.

i. Sketch and find w, b, for the hyperplane $H = \{w^T x = b\}$ found by a hard margin SVM. What are the support vectors and margin width?



ii. How many points can I remove without affecting the resulting hyperplane? There are 10 closed points and 6 open points.

5 SVM Dual

Recall from Stella's SVM lectures that we can formulate the SVM problem in a different way than its primal.

From lecture, we've seen that a convex optimization with a primal of the form

$$\begin{aligned} &\min_x f(x) \\ &\text{s.t. } g(x) \leq 0 \end{aligned}$$

has an associated dual problem, mainly

$$\begin{aligned} &\max_{\alpha} \min_x \mathscr{L}(x, \alpha) \\ &\text{s.t. } \alpha \geq 0 \end{aligned}$$

where $\mathscr{L}(x, \alpha)$ is known as the Lagrangian of the primal. These optimization problems are "equivalent" in the sense that they share the same optimal value and an optimal solution of the primal can be recovered easily from the dual and vice versa.

For the dual to be optimal, recall the KKT conditions:

1. Stationary: $\frac{d\mathscr{L}}{dx} = \frac{df(x)}{dx} + \alpha^T \frac{dg(x)}{dx} = 0$
2. Primal feasibility: $g(x) \leq 0$
3. Dual feasibility: $\alpha \geq 0$
4. Complementary slackness: $\alpha_i g(x)_i = 0$

Using this knowledge, you will now derive dual SVM.

(a) Recall that the primal problem of SVM is of the form

$$\begin{aligned} \max_{w,t\xi} \varepsilon(w,t,\xi) = \max_{w,t,\xi} \frac{1}{2}||w||_2^2 + C\sum_{i=1}^{n} \xi_i \\ \text{s.t. } y_i(w \cdot x_i - t) \geq 1 - \xi_i,\ \xi_i \geq 0 \end{aligned}$$

What is the lagrangian $\mathscr{L}(w,t,\xi,\alpha,\beta)$ for the primal? Use α_i for the $y_i(w \cdot x_i - t) \geq 1 - \xi_i$ constraint and β_i for the $\xi_i \geq 0$ constraint.

(b) Using KKT conditions, derive the value of w in terms of dual variables α_i at the optimal point. What is significant about this value of w? Furthermore, prove that at optimum for the dual, $0 \leq \alpha_i \leq C$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$.

(c) Write the dual problem for SVM in terms of the dual variables $\alpha = \begin{bmatrix}\alpha_1 & \dots & \alpha_n\end{bmatrix}^T$ and matrix Q, where $Q_{ij} = y_i y_j (x_i^T x_j)$.

(d) Prove the following (HINT: use the "Complementary slackness" KKT condition):

(a) $\alpha_i = 0$ means x_i is on or outside of the margin.

(b) $\alpha_i = C$ means x_i violates the margin.

(c) $0 < \alpha_i < C$ means x_i is a support vector.

6 Logistic Regression

Consider the log-likelihood function for logistic regression

$$\ell(\theta) = \sum_{i=1}^{m} y^{(i)} \ln h(x^{(i)}) + (1 - y^{(i)}) \ln(1 - h(x^{(i)}))$$

Find the Hessian H of this function , and show that for every z, it holds true that

$$z^\top H z \leq 0$$

(Hint: You might want to start by showing that $\sum_i \sum_j z_i x_i x_j z_j = (x^\top z)^2$)