10-601 Machine Learning: Homework 6

Due 5 p.m. Friday, April 3, 2015

Instructions

- **Late homework policy:** Homework is worth full credit if submitted before the due date, half credit during the next 48 hours, and zero credit after that. You *must* turn in at least $n-1$ of the n homeworks to pass the class, even if for zero credit.
- **Collaboration policy:** Homeworks must be done individually, except where otherwise noted in the assignments. "Individually" means each student must hand in their own answers, and each student must write and use their own code in the programming parts of the assignment. It is acceptable for students to collaborate in figuring out answers and to help each other solve the problems, though you must in the end write up your own solutions individually, and you must list the names of students you discussed this with. We will be assuming that, as participants in a graduate course, you will be taking the responsibility to make sure you personally understand the solution to any work arising from such collaboration.
- **Online submission:** You must submit your solutions online on autolab. We recommend that you use LaTeX to type your solutions to the written questions, but we will accept scanned solutions as well. On the Homework 6 autolab page, you can download the template, which is a tar archive containing a blank placeholder pdf for the written questions. Replace each pdf file with one that contains your solutions to the written questions. When you are ready to submit, create a new tar archive of the top-level directory and submit your archived solutions online by clicking the "Submit File" button. You should submit a single tar archive identical to the template, except with the blank pdfs replaced by your solutions for the written questions. You are free to submit as many times as you like. ***DO NOT*** change the name of any of the files or folders in the submission template. In other words, your submitted files should have exactly the same names as those in the submission template. Do not modify the directory structure.

Problem 1: Expectation Maximization

Mixture models are helpful for modeling unknown subpopulations in data. If we have a collection of data points $X = \{X_1, ..., X_n\}$, where each X_i is drawn from one of K possible distributions, we can introduce a discrete-valued random variable $Z_i \in \{1, ..., K\}$ that indicates which distribution X_i is drawn from. In lecture, we discussed Gaussian mixture models, where each sample X_i is drawn from a Gaussian distribution according to the value of its mixture component. This question deals with the *categorical* mixture model, where each observation X_i is a discrete value drawn from a categorical distribution, rather than a continuous value drawn from a Gaussian distribution. The parameter for a categorical distribution is a K-dimensional vector π that lists the probability of each of K possible values (therefore, $\sum_k \pi_k = 1$). For example, suppose our categorical mixture model has 3 underlying distributions. Then, each Z_i could take on one of three values, $\{1, 2, 3\}$, with respective probabilities $\{\pi_1, \pi_2, \pi_3\}$; equivalently, $Z_i \sim \text{Categorical}(\pi)$, where $\pi = [\pi_1, \pi_2, \pi_3]^T \in \mathbb{R}^3$. The observation X_i is then generated from another categorical distribution, depending on the value of Z_i. Equation (1) summarizes the generative process for a categorical mixture model.

$$\begin{aligned} Z_i &\sim \text{Categorical}(\pi) \\ X_i &\sim \text{Categorical}(\theta_{Z_i}) \end{aligned} \tag{1}$$

For this model, where we observe X but not Z, we want to learn the parameters of the K categorical components $\Theta = \{\pi, \theta_1, ..., \theta_K\}$, where each $\theta_k \in \mathbb{R}^M$ is the parameter for the categorical distribution associated with the k-th mixture component (this means that each X_i can take on one of M possible values). We can use the EM algorithm to accomplish this.

A note on notation and a hint: it is helpful to use indicator variables when working with categorical distributions. The indicator function $\mathbb{1}\{x = j\}$ has value 1 when $x = j$ and 0 otherwise. With this notation, we can express the probability that a random variable drawn from a categorical distribution (e.g., $Y \sim$ Categorical(ϕ), where $\phi \in \mathbb{R}^N$) takes on a particular value as

$$P(Y) = \prod_{i=1}^{N} \phi_i^{\mathbb{1}\{Y=i\}}.$$

(a) **[6 points]** What is the joint distribution $P(X, Z; \Theta)$?

Solution:

$$\begin{aligned} P(X, Z; \Theta) &= \prod_{i=1}^{n} P(Z_i; \Theta) P(X_i | Z_i; \Theta) \\ &= \prod_{i=1}^{n} \prod_{k=1}^{K} \left[\pi_k \prod_{j=1}^{M} \theta_{k,j}^{\mathbb{1}\{x_i=j\}} \right]^{\mathbb{1}\{z_i=k\}} \end{aligned} \tag{2}$$

(b) **[6 points]** What is the posterior distribution of the latent variables, $P(Z|X; \Theta)$?

Solution:

$$\begin{aligned} P(Z|X; \Theta) &= \prod_{i=1}^{n} P(Z_i | X_i; \Theta) \\ &= \prod_{i=1}^{n} \frac{P(X_i|Z_i; \Theta) P(Z_i; \Theta)}{P(X_i; \Theta)} \\ &= \prod_{i=1}^{n} \frac{P(X_i|Z_i; \Theta) P(Z_i; \Theta)}{\sum_{\ell=1}^{K} P(z_i = \ell; \Theta) P(X_i | z_i = \ell; \Theta)} \\ &= \prod_{i=1}^{n} \frac{\prod_{k=1}^{K} \left[\pi_k \prod_{j=1}^{M} \theta_{k,j}^{\mathbb{1}\{x_i=j\}} \right]^{\mathbb{1}\{z_i=k\}}}{\sum_{\ell=1}^{K} \pi_\ell \prod_{j=1}^{M} \theta_{\ell,j}^{\mathbb{1}\{x_i=j\}}} \end{aligned} \tag{3}$$

(c) **[8 points]** What is the expectation of the log-likelihood, $Q(\Theta'|\Theta) := \mathbb{E}_{Z|X;\Theta} [\log P(X, Z; \Theta')]$?

Solution:

$$
\begin{aligned}
Q(\Theta'|\Theta) &= \mathbb{E}_{Z|X;\Theta}\left[\log P(X, Z; \Theta')\right] \\
&= \mathbb{E}_{Z|X;\Theta}\left[\log \prod_{i=1}^{n} P(X_i, Z_i; \Theta')\right] \\
&= \mathbb{E}_{Z|X;\Theta}\left[\log \prod_{i=1}^{n} P(Z_i; \Theta') P(X_i|Z_i; \Theta)\right] \\
&= \mathbb{E}_{Z|X;\Theta}\left[\log \prod_{i=1}^{n} \prod_{k=1}^{K} \left[\pi_k \prod_{j=1}^{M} \theta_{k,j}^{\mathbb{1}\{x_i=j\}}\right]^{\mathbb{1}\{z_i=k\}}\right] \\
&= \mathbb{E}_{Z|X;\Theta}\left[\sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}\{z_i = k\} \left(\log \pi_k + \sum_{j=1}^{M} \mathbb{1}\{x_i = j\} \log \theta_{k,j}\right)\right] \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{E}_{Z|X;\Theta}\left[\mathbb{1}\{z_i = k\}\right] \left(\log \pi_k + \sum_{j=1}^{M} \mathbb{1}\{x_i = j\} \log \theta_{k,j}\right)
\end{aligned} \tag{4}
$$

We'll now take a short detour to calculate the expectation of the latent variables, given the observations. First, note that the expectation of an indicator is just the probability of the event within the indicator.

$$
\begin{aligned}
\mathbb{E}_{Z|X;\Theta}\left[\mathbb{1}\{z_i = k\}\right] &= P(z_i = k|X_i; \Theta) \\
&= \frac{P(z_i = k) P(X_i|z_i = k)}{P(X_i)} \\
&= \frac{P(z_i = k) P(X_i|z_i = k)}{\sum_{\ell=1}^{K} P(z_i = \ell) P(X_i|z_i = \ell)} \\
&= \frac{P(z_i = k) \prod_{j=1}^{M} \theta_{k,j}^{\mathbb{1}\{x_i=j\}}}{\sum_{\ell=1}^{K} P(z_i = \ell) \prod_{j=1}^{M} \theta_{\ell,j}^{\mathbb{1}\{x_i=j\}}} \\
&= \frac{\pi_k \prod_{j=1}^{M} \theta_{k,j}^{\mathbb{1}\{x_i=j\}}}{\sum_{\ell=1}^{K} \pi_\ell \prod_{j=1}^{M} \theta_{\ell,j}^{\mathbb{1}\{x_i=j\}}} \\
&=: \gamma(z_{ik})
\end{aligned} \tag{5}
$$

Note that these quantities, which we'll call $\gamma(z_{ik})$, are based on our current parameter estimate Θ, so we can calculate them. Now we can finish the derivation of $Q(\Theta'|\Theta)$ from (4), by substituting in these $\gamma(z_{ik})$ values for the expectation of the latent variables.

$$
Q(\Theta'|\Theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma(z_{ik}) \left(\log \pi_k + \sum_{j=1}^{M} \mathbb{1}\{x_i = j\} \log \theta_{k,j}\right) \tag{6}
$$

(d) **[Extra Credit, 5 points]** What is the update step for Θ? That is, what Θ maximizes $Q(\Theta'|\Theta)$? (Remember that Θ includes the categorical parameter π for Z and the categorical parameters $\theta_1, ..., \theta_K$ for X, from the generative process described in (1). Make sure your solution enforces the constraint that parameters to categorical distributions must sum to 1—Lagrange multipliers are a great way to solve constrained optimization problems.)

Solution: To find the updates for the parameters, we solve the following constrained optimization problem:

$$\begin{aligned} &\max_{\pi,\theta_{1:K}} Q(\Theta'|\Theta) \\ \text{subject to } &\sum_{k=1}^{K} \pi_k = 1 \\ &\sum_{j=1}^{M} \theta_{k,j} = 1, \ \forall k = 1, ..., K \end{aligned}$$

We introduce a Lagrange multiplier for each constraint and form the Lagrangian:

$$\mathcal{L} = Q(\Theta'|\Theta) + \lambda_\pi \left(1 - \sum_{k=1}^{K} \pi_k\right) + \sum_{k=1}^{K} \lambda_{\theta_k} \left(1 - \sum_{j=1}^{M} \theta_{k,j}\right) \tag{7}$$

Now we differentiate the Lagrangian with respect to each parameter we want to optimize, set the derivative to zero, and solve for the optimal value.

We'll start with optimizing for π, the parameter to the categorical distribution for the latent variables Z.

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \pi_k} &= \sum_{i=1}^{n} \gamma(z_{ik}) \cdot \frac{1}{\pi_k} - \lambda_\pi \overset{\text{set}}{=} 0 \\ \pi_k^\star &= \frac{1}{\lambda_\pi} \sum_{i=1}^{n} \gamma(z_{ik}) \end{aligned} \tag{8}$$

We can find the value of the Lagrange multiplier λ_π by enforcing the constraint:

$$\begin{aligned} 1 &= \sum_{k=1}^{K} \pi_k^\star \\ 1 &= \sum_{k=1}^{K} \frac{1}{\lambda_\pi} \sum_{i=1}^{n} \\ \lambda_\pi &= \sum_{k=1}^{K} \sum_{i=1}^{n} \gamma(z_{ik}) = n \end{aligned}$$

Substituting this value for λ_π back into the expression from (8) gives the answer:

$$\pi_k^\star = \frac{1}{n} \sum_{i=1}^{n} \gamma(z_{ik}).$$

The process for optimizing the observation parameters $\theta_{k,j}$ is similar. First, we differentiate the Lagrangian from (7) with respect to $\theta_{k,j}$, set to zero, and solve.

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \theta_{k,j}} &= \sum_{i=1}^{n} \gamma(z_{ik}) \mathbb{1}\{x_i = j\} \cdot \frac{1}{\theta_{k,j}} - \lambda_{\theta_k} \overset{\text{set}}{=} 0 \\ \theta_{k,j}^\star &= \frac{1}{\lambda_{\theta_k}} \sum_{i=1}^{n} \gamma(z_{ik}) \mathbb{1}\{x_i = j\} \end{aligned} \tag{9}$$

Again, we can find the value of the Lagrange multiplier λ_{θ_k} by enforcing the summation constraint:

$$\begin{aligned} 1 &= \sum_{j=1}^{M} \theta^{\star}_{k,j} \\ 1 &= \sum_{j=1}^{M} \frac{1}{\lambda_{\theta_k}} \sum_{i=1}^{n} \gamma(z_{ik}) \mathbb{1}\{x_i = j\} \\ \lambda_{\theta_k} &= \sum_{j=1}^{M} \sum_{i=1}^{n} \gamma(z_{ik}) \mathbb{1}\{x_i = j\} \end{aligned}$$

Substituting this value for λ_{θ_k} back into the expression from (9) gives the answer:

$$\begin{aligned} \theta^{\star}_{k,j} &= \frac{\sum_{i=1}^{n} \gamma(z_{ik}) \mathbb{1}\{x_i = j\}}{\sum_{\ell=1}^{M} \sum_{i=1}^{n} \gamma(z_{ik}) \mathbb{1}\{x_i = \ell\}} \\ &= \sum_{i=1}^{n} \frac{\gamma(z_{ik}) \mathbb{1}\{x_i = j\}}{\sum_{\ell=1}^{M} \gamma(z_{ik}) \mathbb{1}\{x_i = \ell\}} \end{aligned}$$

Problem 2: AdaBoost

Consider the following dataset, plotted in Figure 1:

$$\begin{aligned} X_1 = (-1, 0, +), X_2 = (-0.5, 0.5, +), X_3 = (0, 1, -), X_4 = (0.5, 1, -), \\ X_5 = (1, 0, +), X_6 = (1, -1, +), X_7 = (0, -1, -), X_8 = (0, 0, -). \end{aligned}$$

In this problem, you'll run through $T = 3$ iterations of AdaBoost with decision stumps (axis-aligned half-planes) as weak learners.

(a) **[20 points]** For each iteration $t = 1, 2, 3$, compute $\epsilon_t, \alpha_t, Z_t, D_t(i) \, \forall i$ (in Table 1) and draw the decision stump (on Figure 1). Recall that Z_t is the normalization factor to ensure that the weights D_t sum to one.

Solution: See Figure 1 and Table 1.

(b) **[5 points]** What is the training error of AdaBoost? Give a one-sentence reason for why AdaBoost outperforms a single decision stump.

Solution: The final AdaBoost classifier here has zero training error. AdaBoost outperforms a single decision stump because it combines several weak learners to produce a nonlinear decision boundary.

Table 1: Values of AdaBoost parameters at each timestep.

t	ϵ_t	α_t	Z_t	$D_t(1)$	$D_t(2)$	$D_t(3)$	$D_t(4)$	$D_t(5)$	$D_t(6)$	$D_t(7)$	$D_t(8)$
1	0.250	0.549	0.866	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
2	0.167	0.805	0.745	0.083	0.083	0.083	0.083	0.250	0.250	0.083	0.083
3	0.100	1.099	0.600	0.250	0.250	0.050	0.050	0.150	0.150	0.050	0.050

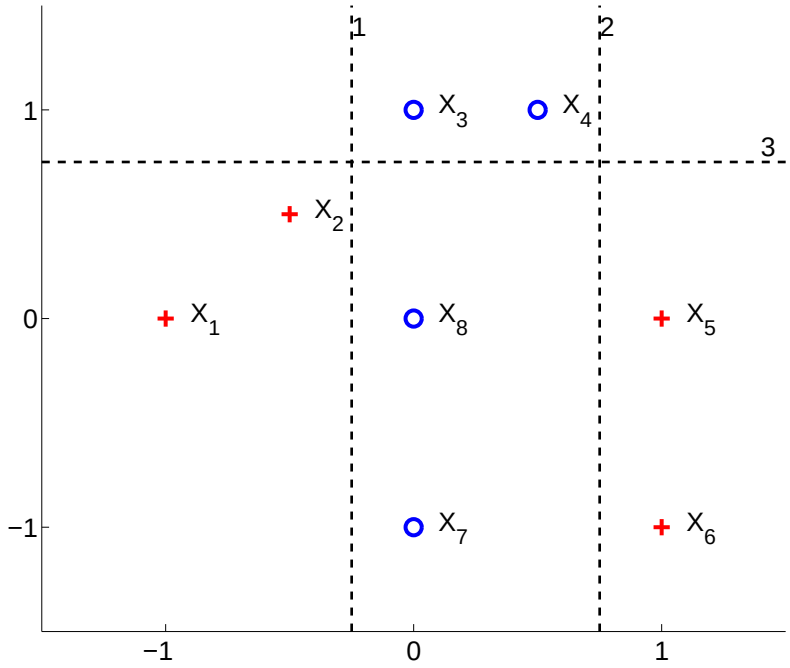

Figure 1: A small dataset, for binary classification with AdaBoost.

Problem 3: Perceptron

Consider running the Perceptron algorithm on some sequence of examples S (an example is a data point and its label). Let S' be the same set of examples as S, but presented in a different order.

(a) **[4 points]** Does the Perceptron algorithm necessarily make the same number of mistakes on S as it does on S'?

Solution: No.

(b) **[8 points]** If so, why? If not, show such an S and S' where the Perceptron algorithm makes a different number of mistakes on S' than it does on S.

Solution: Consider the set points plotted in Figures 2 and 3. The ordering in Figure 2, $\{1, 2, 3, 4, 5, 6\}$, results in 3 mistakes. The ordering in Figure 3, $\{1, 4, 5, 2, 3, 6\}$, results in 2 mistakes.

(c) **[8 points]** We know that in $\mathbb{R}^d$ we can shatter $d+1$ points with linear separators, but not $d+2$ points (because the VC-dimension of linear separators is $d+1$). But what if we require that the points be separated by margin γ? Show that you can have at most $(R/\gamma)^2$ points inside the ball of radius R that can be "shattered at margin γ," meaning that every labeling is achievable by a separator of margin γ. **Hint:** Suppose for contradiction you had $M = (R/\gamma)^2 + 1$ such points. What would happen if you gave them in some order to the Perceptron algorithm, with labels exactly opposite of what Perceptron predicts?

Solution: Suppose, for a contradiction, that there are $(R/\gamma)^2 + 1$ such points, presented in an order such that Perceptron gets them all wrong. That is, the algorithm made more than $(R/\gamma)^2$ mistakes. Therefore, the dataset must not have had margin γ with all points in a ball of radius R (contrapositive of theorem that, if a dataset has margin γ and all points inside a ball of radius R, then Perceptron makes at most $(R/\gamma)^2$ mistakes.

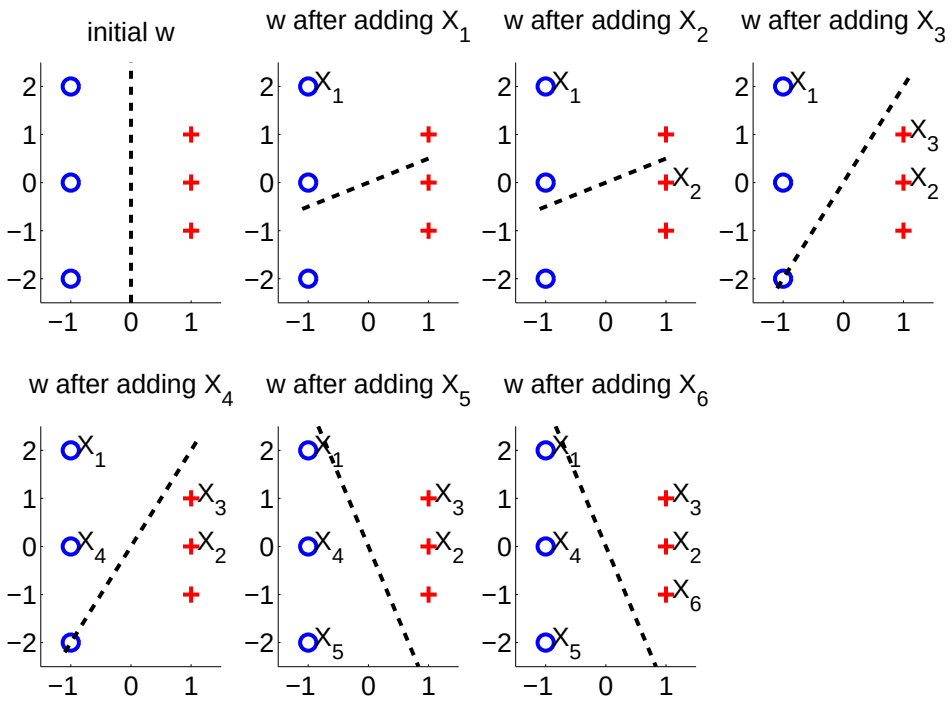

Figure 2: With this ordering of points, Perceptron makes 3 mistakes (on X_1, X_3, X_5).

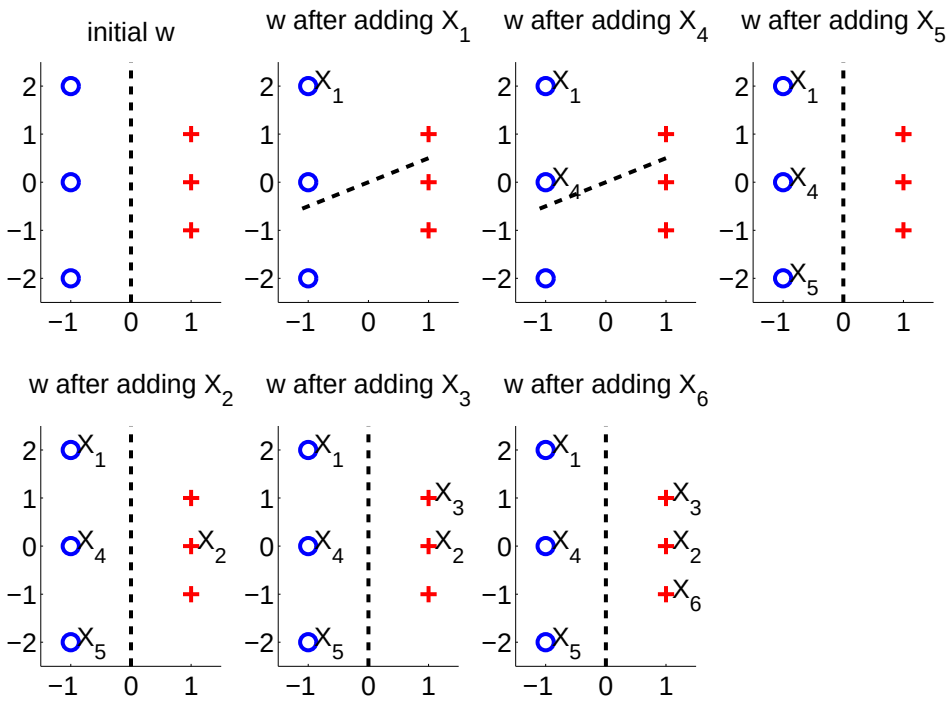

Figure 3: With this ordering of points, Perceptron makes 2 mistakes (on X_1, X_5).

Problem 4: Kernels

4.1: A proposed kernel

Consider the following kernel function:

$$K(x, x') = \begin{cases} 1, & \text{if } x = x' \\ 0, & \text{otherwise} \end{cases}$$

(a) **[8 points]** Prove this is a legal kernel. That is, describe an implicit mapping $\Phi : X \to \mathbb{R}^m$ such that $K(x, x') = \Phi(x) \cdot \Phi(x')$. (You may assume the instance space X is finite.)

Solution: For any collection of points $x_1 \ldots, x_m$, the gram matrix is given by

$$G = \begin{bmatrix} K(x_1, x_1) & \ldots & K(x_1, x_m) \\ \vdots & \ddots & \vdots \\ K(x_m, x_1) & \ldots & K(x_m, x_m) \end{bmatrix} = I,$$

where I denotes the identity matix in $\mathbb{R}^m$. Since this matrix is PSD, it follows by Mercer's theorem that K is a valid kernel.

Moreover, for any fixed collection of points $x_1, \ldots, x_m$, the map $\Phi(x_i) = e_i$, where e_i is the i^{th} standard basis vector in $\mathbb{R}^m$, satisfies $K(x_i, x_j) = \Phi(x_i)^\top \Phi(x_j)$.

(b) **[6 points]** In this kernel space, any labeling of points in X will be linearly separable. Justify this claim.

Solution: Fix any set of points $x_1, \ldots, x_m$, any labeling $y_1, \ldots, y_m \in \{\pm 1\}$ and let Φ be the feature map from part (a). Then let $w \in \mathbb{R}^m$ be the vector with components $w_i = y_i$. Then for each x_i, we have that

$$w^\top \Phi(x_i) = w^\top e_i = w_i = y_i.$$

From this, it follows that if $y_i = 1$ then $\Phi(x_i)$ is on the positive side of the linear separator given by $h(x) = w^\top x$, and if $y_i = -1$ then $\Phi(x_i)$ is on the negative side. Since this construction works for any labels and any points, this shows that under this kernel, every data set is linearly separable in the feature space.

(c) **[6 points]** Since all labelings are linearly separable, this kernel seems perfect for learning any target function. Why is this actually a bad idea?

Solution: This is a dangerous kernel to use, since it is very likely that we will overfit the data.

4.2: Composition of kernels

It is possible to use previously-defined kernels to construct new ones. For this problem, you will prove why the following proposed functions are or are not valid kernels The notation $< x, z >$ indicates the dot product $x^T z$).

(a) **[3 points]** $K(x, z) = 5 < x, z >$

Solution: Let $\Phi(x) = \sqrt{5}x$. Then $K(x, z) = 5x^\top z = (\sqrt{5}x)^\top(\sqrt{5}z) = \Phi(x)^\top \Phi(y)$. Therefore, K is a valid kernel, since it corresponds to the feature map Φ.

(b) **[4 points]** $K(x, z) =< x, z >^3 +(< x, z > +1)^2$

Solution: We can rewrite $K(x, z) = (x^\top z)^3 + (x^\top z)^2 + 2(x^\top z) + 1$. We can see that the functions $k_3(x, z) = (x^\top z)^3$, $k_2(x, y) = (x^\top z)^2$, and $k_1(x, z) = (x^\top z)$ are all kernels from the product closure property. The function $k_0(x, z) = 1$ is a kernel with feature map $\Phi(x) = 1$. Therefore, K is a valid kernel since it is a linear combination of $k_0, \ldots, k_3$ with positive coefficients.

(c) **[8 points]** $K(x, z) =< x, z >^2 + \exp(-||x||^2) \exp(-||z||^2)$

Solution: Again, the function $k_2(x, z) = (x^\top z)^2$ is a kernel. Letting $\Phi(x) = \exp(-\|x\|^2)$, we see that the second term is the kernel with feature map Φ. It follows that K is a valid kernel, since it is a linear combination of two kernels with positive coefficients.

Problem 5: Extra Credit

Support Vector Machines

Suppose we have a set $S = \{(x_1, y_1), ..., (x_m, y_m)\}$ of labeled examples in $\mathbb{R}^n$, and assume $||x_i|| = 1, \forall i$. It is NP-hard to find a linear separator that minimizes the number of points misclassified, so learning algorithms optimize other related quantities that are easier to solve. SVM solves the optimization problem

$$\begin{aligned} \min_{w} \quad & ||w||_2^2 + C\sum_{i=1}^{m} \xi_i \\ \text{subject to} \quad & y_i(w^T x_i) \geq 1 - \xi_i \quad \forall i \\ & \xi_i \geq 0 \quad \forall i \end{aligned} \tag{10}$$

(a) **[5 points]** Suppose that S has the property that the total distance one would need to move the points in order to make them separable by margin γ is d_γ. (By "total distance" we mean $\sum_i d_i$ where d_i is the distance that point x_i is moved. By "separable by margin γ" we mean that, for some hyperplane through the origin, all points are on the correct side and at distance at least γ from it.) Show that, for an appropriate value of C, the number of misclassifications made on S by the separator produced by SVM is at most $\frac{1}{2} + d_\gamma/\gamma$.
Hint: We can find the value of d_γ with the following optimization problem:

$$\begin{aligned} \min_{w, \tilde{\gamma}_i, \tilde{\gamma}} \quad & \sum_{i=1}^{m} \tilde{\gamma}_i \\ \text{subject to} \quad & ||w|| = 1 \\ & y_i(w^T x_i) \geq \tilde{\gamma} - \tilde{\gamma}_i \quad \forall i \end{aligned} \tag{11}$$

where γ_i indicates the distance between point x_i and the appropriate margin.

Solution: Let w, γ, γ_i be the optimum for (11). Then, $d_\gamma = \sum_{i=1}^{m} \gamma_i$ and $y_i(1/\gamma)w^T x_i \geq 1 - \gamma_i/\gamma$.

Let $\tilde{w} = (1/\gamma)w$; then, $y_i \tilde{w}^T x_i \geq 1 - \gamma_i/\gamma$.

Let w_S be the optimum from the SVM formulation, with value cost(w_S). Since w_S is the optimum, then, by definition, cost$(w_S) \leq$ cost$(\tilde{w})$. Furthermore, cost$(\tilde{w}) \leq 1/\gamma^2 + Cd_\gamma/\gamma$. Therefore, the number of misclassifications w_S makes on S is at most $(1/C)$cost$(\tilde{w}) \leq \frac{1}{C}\frac{1}{\gamma} + \frac{d_\gamma}{\gamma}$. When $C = 2/\gamma^2$, this upper bound on the number of misclassifications is $\frac{1}{2} + d_\gamma/\gamma$.

Boosting

Suppose that, instead of computing $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$, we instead fixed $\alpha > 0$ ahead of time and set $\alpha_t = \alpha$ for each iteration t in this Boost(α) algorithm. Assume that on every round t of boosting, we know that ϵ_t will be at most $1/2 - \gamma$, for some $\gamma > 0$. We choose to set $\alpha = \frac{1}{2}\ln\left(\frac{1+2\gamma}{1-2\gamma}\right)$.

(a) **[5 points]** Show how to modify the training error analysis of AdaBoost to derive an upper bound on the training error of the final hypothesis H produced by Boost(α) after T rounds. Your bound should be in terms of γ and T only (it should not depend on α,ϵ_t, etc.). **Hint:** Begin by proving, as we did in class, the bound on the error $\text{err}_S(H_{\text{final}})$ in terms of the normalization factors Z_t.

Solution: We begin by showing, as in lecture, that $\text{err}_S(H_{\text{final}}) \leq \prod_{t=1}^{T} Z_t$:

$$\begin{aligned}
\text{err}_S(H_{\text{final}}) &= \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\left\{y_i \neq H_{\text{final}}(x_i)\right\} \\
&= \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\left\{y_i f(x_i) \leq 0\right\} \\
&\leq \frac{1}{m}\sum_{i=1}^{m} \exp(-y_i f(x_i)) \qquad (12) \\
&= \sum_{i=1}^{m} D_{T+1}(i) \prod_{t=1}^{T} Z_t \\
&= \prod_{t=1}^{T} Z_t
\end{aligned}$$

Now, let's bound Z_t:

$$\begin{aligned}
Z_t &= (1-\epsilon_t)\exp(-\alpha) + \epsilon_t \exp(\alpha) \\
&= \epsilon_t(\exp(\alpha) - \exp(-\alpha)) + \exp(-\alpha) \\
&\leq \left(\frac{1}{2} - \gamma\right)(\exp(\alpha) - \exp(-\alpha)) + \exp(-\alpha) \\
&= \left(\frac{1}{2} - \gamma\right)\exp(\alpha) + \left(\frac{1}{2} + \gamma\right)\exp(-\alpha) \\
&= \left(\frac{1}{2} - \gamma\right)\left(\frac{1+2\gamma}{1-2\gamma}\right)^{1/2} + \left(\frac{1}{2} + \gamma\right)\left(\frac{1+2\gamma}{1-2\gamma}\right)^{-1/2} \\
&= \frac{1}{2}(1-2\gamma)(1+2\gamma)^{1/2}(1-2\gamma)^{-1/2} + \frac{1}{2}(1+2\gamma)(1+2\gamma)^{-1/2}(1-2\gamma)^{1/2} \\
&= \left[(1+2\gamma)(1-2\gamma)\right]^{1/2} \\
&= (1-4\gamma^2)^{1/2}
\end{aligned}$$

We can substitute this bound on Z_t into the result from (12) to bound the final training error:

$$\begin{aligned}
\text{err}_S(H_{\text{final}}) &\leq \prod_{t=1}^{T} Z_t \\
&\leq \prod_{t=1}^{T} (1-4\gamma^2)^{1/2} \\
&= (1-4\gamma^2)^{T/2} \\
&\leq \exp(-2T\gamma^2)
\end{aligned}$$

(b) **[5 points]** Use the previous result to show that the final hypothesis H will be consistent with m training examples (i.e., have zero training error) after T rounds if $T > \frac{\ln m}{2\gamma^2}$.

Solution: Suppose we have m training samples $\{(x_1, y_1), ..., (x_m, y_m)\}$. Then the training error will be given by $\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\left\{y_i \neq f(x_i)\right\}$, where $f(x)$ is our final classifier. If this value is less than $\frac{1}{m}$, then the training error

is 0.

$$\begin{aligned} \text{err}_S(H_{\text{final}}) &< \frac{1}{m} \\ \exp(-2T\gamma^2) &< \frac{1}{m} \\ T &> \frac{\ln m}{2\gamma^2} \end{aligned}$$

(c) **[5 points]** Assume that the weak learning algorithm generates hypotheses h_t which belong to a finite class $\mathcal{H}$. Use error bounds from the class to show that if we choose T as in part (b), then with probability $1-\delta$, the generalization error of H is at most $\frac{T \ln |\mathcal{H}| + \ln(1/\delta)}{m}$.

Solution: Since we choose T such that the final hypothesis is consistent with the training data and $|\mathcal{H}| < \infty$, we are in the realizable case for finite hypothesis spaces. The effective hypothesis space is of size $|\mathcal{H}|^T$, since our final hypothesis H is composed of T weak learners, each from a hypothesis space of size $|\mathcal{H}|$. Thus, we use the statistical learning theory bound:

$$\text{err}_D(h_{\text{final}}) \leq \frac{1}{m}\left(\ln |\mathcal{H}|^T + \ln(1/\delta)\right) = \frac{1}{m}\left(T \ln |\mathcal{H}| + \ln(1/\delta)\right)$$