Midterm Exam

Prof. David S. Rosenberg

April 11, 2015

1 True / False Questions

1. (**True or False**, 1 pt) When using (unregularized) linear regression, adding new features always improves the performance on training data, or at least never make it worse.

2. (**True or False**, 1 pt)When using a (unregularized) linear regression, adding new features always improves the performance on test data, or at least never make it worse.

3. (**True or False**, 1 pt) Overfitting is more likely when the set of training data is small.

4. (**True or False**, 1 pt) Overfitting is more likely when the hypothesis space is small.

5. (**True or False**, 1 pt) Approximation error decreases to zero as the amount of training data goes to infinity.

6. (**True or False**, 1 pt) If the empirical risk function is not convex, increasing training data may not decrease the expected estimation error.

7. (**True or False**, 1 pt) If a decision tree is trained on data for which two features are exactly equal, the resulting tree will be the same whether or not we remove one of those two features.

8. (**True or False**, 1 pt) Adaboost with decision stumps will eventually reach zero training error, provided we run enough rounds of boosting.

2 Short Answer

1. (1 pt) Which of the following loss functions may lead to support vectors: hinge loss, logistic loss, square loss.

2. (2 pts) Show that the following kernel function is a Mercer kernel (i.e. it represents an inner product):

$$k(x, y) = \frac{x^T y}{\|x\|\|y\|},$$

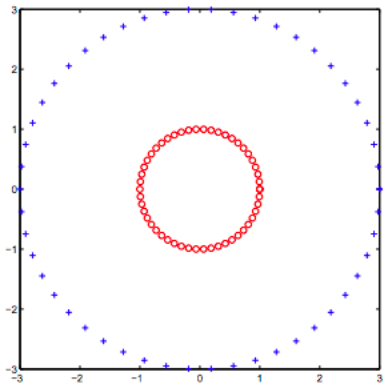


Figure 2.1: For a short-answer problem.

where $x, y \in \mathbf{R}^d$.

For $\phi(x) = \frac{x}{\|x\|}$, we have

$$k(x, y) = \langle \phi(x), \phi(y) \rangle .$$

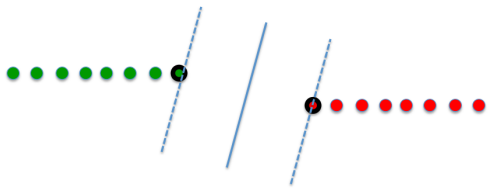
3. (2 pts) Consider the binary classification problem shown in Figure (2.1): Denote the input space by $\mathcal{X} = \{(x_1, x_2) \in \mathbf{R}^2\}$. Give a feature mapping for which a linear classifier could perfectly separate the two classes shown.

4. (2 pts) For the classification problem in Figure (2.1), circle all classifiers that could perfectly separate the classes: **linear SVM, SVM with quadratic kernel, decision stumps (i.e. classification trees with only two leaf nodes), AdaBoost with decision stumps, SVM with radial basis function kernel**.

5. (2 pts) Let $\mathcal{F}_1 = \{\text{binary decision trees of depth 2}\}$. Let $\mathcal{F}_2 = \{\text{all linear classifiers}\}$. Draw a binary classification dataset for which a member of $\mathcal{F}_1$ can perfectly separate the data, while no member of $\mathcal{F}_2$ can. Show the splits and the decision boundary for the tree.

6. (2 pts) Same $\mathcal{F}_1$ and $\mathcal{F}_2$ as in the previous problem. Draw a binary classification dataset for which a member of $\mathcal{F}_2$ can perfectly separate the data, while no member of $\mathcal{F}_1$ can.

7. (2 pts) Suppose we fit a hard-margin SVM to N data points and we have 2 data points "on the margin". If we add a new data point to the training set and refit the SVM, what's the most number of data points that could end up "on the margin". Support your answer (a picture could suffice).



3 Hypothesis Spaces

1. (1 pt) Consider the following two hypothesis spaces:

$$\mathcal{F}_1 = \{f(x) = e^{w_1}x + w_2 x \mid w_1, w_2 \in \mathbf{R}\} \qquad \mathcal{F}_2 = \{f(x) = wx \mid w \in \mathbf{R}\}$$

   Suppose we are selecting hypotheses using empirical risk minimization (without any penalty). Are there any situations in which one of these hypothesis spaces would be preferred to the other? Why?

2. (2 pt) Same question, with the following hypothesis spaces:

$$\mathcal{F}_1 = \{f(x) = e^{w_1}x \mid w_1 \in \mathbf{R}\} \qquad \mathcal{F}_2 = \{f(x) = wx \mid w \in \mathbf{R}\}$$

3. (3 pt) Same question, with the following hypothesis spaces:

$$\mathcal{F}_1 = \{\text{trees of depth at most 2}\} \qquad \mathcal{F}_2 = \{\text{trees with at most 4 leaf nodes}\}$$

4 Kernelizing Ridge Regression

Suppose our input space is $\mathcal{X}$=$\mathbf{R}^d$ and our output space is $\mathcal{Y} = \mathbf{R}$. Let $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be a training set from $\mathcal{X} \times \mathcal{Y}$. We'll use the "design matrix" $X \in \mathbf{R}^{n \times d}$, which has the input vectors as rows:

$$X = \begin{pmatrix} -x_1- \\ \vdots \\ -x_n- \end{pmatrix}.$$

Recall the ridge regression objective function:

$$J(w) = ||Xw - y||^2 + \lambda||w||^2,$$

for $\lambda > 0$.

1. (4 pts) Give a closed form expression for the minimizer of $J(w)$.

$$\begin{aligned} J(w) &= (Xw-y)^T(Xw-y) + \lambda w^T w \\ \partial_w J(w) &= 2X^T(Xw-y) + 2\lambda w \\ \partial_w J(w) = 0 &\iff 2X^TXw + 2\lambda w - 2X^Ty = 0 \\ &\iff (X^TX + \lambda I)w = X^Ty \\ &\iff w = (X^TX + \lambda I)^{-1}X^Ty \end{aligned}$$

2. (2 pts) Show that w^*, the minimizer of $J(w)$, can be written as $w = X^T\alpha$, where $\alpha = \lambda^{-1}(y - Xw)$. [If you don't remember the trick, you may want to leave this problem until the end. The rest of the problems do not depend on this one.]

$$\begin{aligned}(X^TX + \lambda I)w &= X^Ty \\ \lambda w &= X^Ty - X^TXw \\ w &= \frac{1}{\lambda}X^T(y - Xw) \\ w &= X^T\alpha\end{aligned}$$

3. (1 pt) Based on the fact that $w = X^T\alpha$, explain why we say w is "in the span of the data."

   $X^T\alpha$ is a linear combination of the columns of X^T, which contain the input vectors $x_1, \ldots, x_n$:

$$w = \begin{pmatrix} | & \cdots & | \\ x_1 & \cdots & x_n \\ | & \cdots & | \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = \alpha_1 x_1 + \cdots \alpha_n x_n$$

4. (3 pts) Give a kernelized expression for α in terms of the kernel matrix XX^T. (Hint: Plug in $w = X^T\alpha$ to the expression for α.)

$$\begin{aligned}\alpha &= \lambda^{-1}(y - Xw) \\ \lambda\alpha &= y - XX^T\alpha \\ XX^T\alpha + \lambda\alpha &= y \\ \left(XX^T + \lambda I\right)\alpha &= y \\ \alpha &= (\lambda I + XX^T)^{-1}y\end{aligned}$$

5. (2 pts) Give a kernelized expression for the Xw, the predicted values on the training points.

$$\begin{aligned}Xw &= X\left(X^T\alpha\right) \\ &= \left(XX^T\right)(\lambda I + XX^T)^{-1}y\end{aligned}$$

6. (1 pt) Give a kernelized expression for the prediction on new points, stored in the design matrix

$$X_P = \begin{pmatrix} -x_1- \\ \vdots \\ -x_n- \end{pmatrix}.$$

   Predictions are

$$X_Pw = X_PX^T(\lambda I + XX^T)^{-1}y$$

5 Ivanov and Tikhonov Regularization

In lecture there was a claim that the Ivanov and Tikhonov forms of ridge and lasso regression are equivalent. We will now prove a much more general result.

5.1 Tikhonov optimal implies Ivanov optimal

Let $\phi : \mathcal{F} \to \mathbf{R}$ be any performance measure of $f \in \mathcal{F}$ and let $\Omega : \mathcal{F} \to \mathbf{R}$ be any complexity measure. For example, for ridge regression over the linear hypothesis space $\mathcal{F} = \left\{ f_w(x) = w^T x \mid w \in \mathbf{R}^d \right\}$, we would have $\phi(f_w) = \frac{1}{n} \sum_{i=1}^{n} \left(w^T x_i - y_i \right)^2$ and $\Omega(f_w) = w^T w$.

1. (3 pts) Suppose that for some $\lambda > 0$ we have the Tikhonov regularization solution

$$f^* = \arg\min_{f \in \mathcal{F}} \left[\phi(f) + \lambda \Omega(f) \right]. \tag{5.1}$$

Show that f^* is also an Ivanov solution. That is, $\exists r > 0$ such that

$$f^* = \arg\min_{f \in \mathcal{F}} \phi(f) \text{ s.t. } \Omega(f) \leq r. \tag{5.2}$$

(Hint: Start by figuring out what r should be. Then one approach is proof by contradiction: suppose f^* is not the optimum in (5.2) and show that contradicts the fact that f^* solves (5.1).)

Take $r = \Omega(f^*)$. Suppose there exists some g with $\Omega(g) \leq r$ and $\phi(g) < \phi(f^*)$. Then

$$\phi(g) + \Omega(g) < \phi(f^*) + \lambda \Omega(f^*),$$

which means that f^* would not be the minimizer in (5.1).

5.2 Ivanov optimal implies Tikhonov optimal

For the converse, we will restrict our hypothesis space to a parametric set. That is, $\mathcal{F} = \left\{ f_w(x) : \mathcal{X} \to \mathbf{R} \mid w \in \mathbf{R}^d \right\}$. So we will now write ϕ and Ω as functions on $w \in \mathbf{R}^d$.

Let w^* be a solution to the following Ivanov optimization problem:

$$\begin{array}{rl} \text{minimize} & \phi(w) \\ \text{subject to} & \Omega(w) \leq r. \end{array}$$

Assume that strong duality holds for this optimization problem and that the dual solution is attained. Then we will show that there exists a $\lambda \geq 0$ such that $w^* = \arg\min_{w \in \mathbf{R}^d} \left[\phi(w) + \lambda \Omega(w) \right]$.

1. (1 pt) Write the Lagrangian $L(w, \lambda)$ for the Ivanov optimization problem.

The Lagrangian is

$$L(w, \lambda) \;=\; \phi(w) + \lambda \left[\Omega(w) - r \right].$$

2. (2 pts) Write the dual optimization problem in terms of the dual objective function $g(\lambda)$, and give an expression for $g(\lambda)$. [Writing $g(\lambda)$ as an optimization problem is expected - don't try to solve it.]

$$\max_{\lambda \geq 0} g(\lambda)$$

where $g(\lambda) = \min_w \left(\phi(w) + \lambda \left[\Omega(w) - r\right]\right)$.

3. (4 pts) We assumed that the dual solution is attained, so let $\lambda^* = \arg\max_{\lambda \geq 0} g(\lambda)$. We also assumed strong duality, which implies $\phi(w^*) = g(\lambda^*)$. Show that the minimum in the expression for $g(\lambda^*)$ is attained at w^*. [Hint: You can use the same approach we used when we derived the complementary slackness conditions.] Conclude the proof by showing that for the choice of $\lambda = \lambda^*$, we have $w^* = \arg\min_{w \in \mathbf{R}^d} \phi(w) + \lambda\Omega(w)$.

$$\begin{aligned}
\phi(w^*) &= g(\lambda^*) \\
&= \min_w \phi(w) + \lambda^* \left[\Omega(w) - r\right] \\
&\leq \phi(w^*) + \underbrace{\lambda^* \left[\Omega(w^*) - r\right]}_{\leq 0} \\
&\leq \phi(w^*).
\end{aligned}$$

So all the inequalities are equalities. Therefore

$$\begin{aligned}
w^* &= \arg\min_w \phi(w) + \lambda^* \left[\Omega(w) - r\right] \\
&= \arg\min_w \phi(w) + \lambda^* \Omega(w)
\end{aligned}$$

5.3 Ivanov implies Tikhonov for Ridge Regression.

To show that Ivanov implies Tikhonov for the ridge regression problem (square loss with ℓ_2 regularization), we need to demonstrate strong duality and that the dual optimum is attained. Both of these things are implied by Slater's constraint qualifications.

1. (4 pts) Show that the Ivanov form of ridge regression is a convex optimization problem with a a strictly feasible point.

   The Ivanov form of ridge regression is

$$\begin{aligned}
\text{minimize} \quad & \sum_{i=1}^{n} \left(y_i - w^T x_i\right)^2 \\
\text{subject to} \quad & w^T w \leq r.
\end{aligned}$$

   First we show this is a convex optimization problem: The objective is convex in w since we have a nonnegative combination of convex functions $\left(y_i - w^T x_i\right)^2$. Each of these expressions is convex since the square is a convex function, and we're applying it to an affine transformation of w. The constraint function is also a convex function of w. Slater's constraint qualification is satisfied since we can take $w = 0$, so long as $r > 0$. Thus we have strong duality.

6 Square Hinge Loss and Huberized Square Hinge Loss

The squared hinge loss is a margin loss given by

$$\ell(m) = \left[(1-m)_+\right]^2,$$

where $(m)_+ = m1(m > 0)$ is the "positive part" of m.

1. (2 pts) Suppose we have a training set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where $x_i \in \mathcal{X} = \mathbf{R}^d$ and $y_i \in \mathcal{Y} = \{-1, 1\}$, for all $i = 1, \ldots, n$. Consider the linear hypothesis space $\mathcal{F} = \left\{f(x) = w^T x \mid w \in \mathbf{R}^d\right\}$. Write the objective function $J(w)$ for ℓ_2-regularized empirical risk minimization with the square hinge loss over the space $\mathcal{F}$, where $\mathcal{F}$ is parameterized by w.

   $$J(w) = \frac{1}{n}\sum_{i=1}^{n}\left[\left(1 - y_i w^T x_i\right)_+\right]^2 + \lambda\|w\|^2,$$

   for $\lambda > 0$.

2. (2 pts) It turns out that $J(w)$ is differentiable at every w. Give the derivative of $J(w)$.

   $$\frac{\partial J(w)}{\partial w} = 2\lambda w + \frac{1}{n}\sum_{i=1}^{n}\begin{cases} -2\left(1 - y_i w^T x_i\right) y_i x_i & y_i w^T x < 1 \\ 0 & \text{otherwise} \end{cases}$$

3. (3 pts) Give pseudocode or otherwise explain how you would use stochastic gradient descent to find $w^* = \arg\min_w J(w)$. You need to specify your approach to the step size, but you do not have to specify a stopping criterion, though you may if you like.

   - $t = 1$
   - Learning rate η =1
   - $w = 0$
   - Repeat until stopping criterion met:
     - randomly choose (x_i, y_i) from $\mathcal{D}$.
     - $w \leftarrow w - \eta\left(2\lambda w + \begin{cases} -2\left(1 - y_i w^T x_i\right) y_i x_i & y_i w^T x < 1 \\ 0 & \text{otherwise} \end{cases}\right)$
     - $t \leftarrow t + 1$
     - $\eta = 1/t$

4. (2 pts) Justify the claim that the output of SGD can be written in the form:

   $$w = \sum_{i=1}^{n}\beta_i x_i.$$

5. (2 pts) In relation to the SGD algorithm, how would you characterize the x_i's that are support vectors?

6. (2 pts) Show that $J(w)$ is convex. (You are free to cite any of the facts given in Appendix A, or you may use the fact that $J(w)$ is convex if and only its Hessian $\nabla^2 J(w)$ is positive semidefinite.)

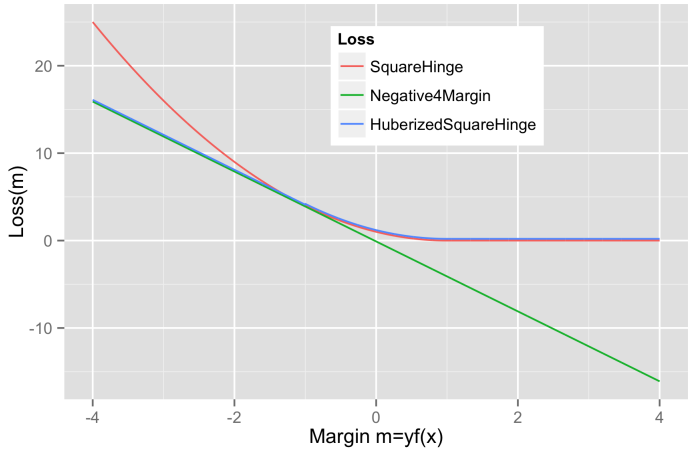
$$\begin{aligned}\nabla^2 J(w) = \frac{\partial^2 J(w)}{\partial w^2} &= 2\lambda I + \frac{1}{n}\sum_{i=1}^{n} \begin{cases} -2\left(-y_i x_i\right)\left(y_i x_i\right)^T & y_i w^T x < 1 \\ 0 & \text{otherwise} \end{cases} \\ &= 2\lambda I + \frac{1}{n}\sum_{i=1}^{n} \begin{cases} 2y_i^2 x_i x_i^T & y_i w^T x < 1 \\ 0 & \text{otherwise} \end{cases}\end{aligned}$$

Note that each $x_i x_i^T$ is a positive semidefinite matrix. So $\nabla^2 J(w)$ is a nonnegative combination of a positive semidefinite matrices, and thus is itself PSD.

7. (2 pts) The "Huberized" square hinge loss is a margin loss given by

$$\ell(m) = \begin{cases} -4m & m < -1 \\ \left[(1-m)_+\right]^2 & \text{otherwise.} \end{cases}$$

The plot below should help explain how this loss relates to the square hinge loss.



- When might you prefer the Huberized square hinge loss to the square hinge loss?

7 Conditional Exponential Distributions

Suppose we want to model the amount of time one will have to wait for a taxi pickup based on the location and the time. The exponential distribution is a natural candidate for this situation. The exponential distribution is a continuous distribution supported on $[0, \infty)$. The set of all exponential probability density functions is given by

$$\text{ExpDists} = \left\{ p_\lambda(y) = \lambda e^{-\lambda y} 1(y \in [0,\infty)) \mid \lambda \in (0,\infty) \right\}.$$

Recall that a family is a **natural exponential family** of continuous distributions on $\mathbf{R}$ with parameter $\theta \in \mathbf{R}$ if its densities can be written as

$$p_\theta(y) = \frac{1}{Z(\theta)} h(y) \exp\left[\theta y\right],$$

where $Z(\theta) = \int h(y) \exp\left[\theta y\right] \, dy$ is the **partition function**. θ is called the **natural parameter**, and the **natural parameter space** Θ consists of all θ for which $Z(\theta) < \infty$. $h(y)$ is called the **base measure.**

1. (4 pts) Write the family of exponential distributions as a natural exponential family. Give expressions for the base measure and the partition function. Identify the natural parameter space.

   Let $h(y) = 1(x \geq 0)$. Then

   $$Z(\theta) = \int_{[0,\infty)} e^{\theta y} \, dy \quad = \quad \frac{1}{\theta} e^{\theta \infty} - \frac{1}{\theta} = \begin{cases} -\frac{1}{\theta} & \theta < 0 \\ \infty & \text{otherwise} \end{cases}$$

   So the natural parameter space is $\Theta = (-\infty, 0)$, and

   $$p_\theta(y) = -\theta 1(y \geq 0) e^{\theta y}.$$

2. (3 pts) Let $x \in \mathbf{R}^d$ represent the input features from which we want to predict an exponential distribution. We will use a generalized linear model (GLM) approach. Suggest a reasonable function ψ to map $w^T x$ to a value in the natural parameter space Θ. Then write an expression for $p_w(y \mid x)$, the predicted probability density function conditioned on x.

   Since $\Theta = (-\infty, 0)$, we'll use $\psi(\eta) = -e^\eta$. So

   $$p_w(y \mid x) = e^{w^T x} 1(y \geq 0) e^{-\exp\left(w^T x\right) y}$$

3. (3 pts) Suppose we have a data set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where $x_i \in \mathbf{R}^d$ and $y_i \in [0, \infty)$ for $i = 1, \ldots, n$. Give the optimization problem you would solve to fit the GLM we have been discussing to training data $\mathcal{D}$.

   We'll find the ERM for the loss function $-\log p(y \mid x)$. We have

   $$\log p_w(y \mid x) = w^T x - \exp\left(w^T x\right) y,$$

   so we'll take the objective function to be

   $$\begin{aligned} J(w) &= -\frac{1}{n} \sum_{i=1}^n \log p_w(y_i \mid x_i) \\ &= \frac{1}{n} \sum_{i=1}^n \exp\left(w^T x_i\right) y_i - w^T x_i \end{aligned}$$

   We'd look for $w^* = \arg\min_{w \in \mathbf{R}^d} J(w)$.

4. (5 pts) Suppose we think that a linear function of x doesn't extract enough information, and we'd like to use a more expressive model. For full credit, explain how you would use gradient boosting in this situation. For partial credit, present another reasonable approach to this problem.

   As before, we need a function $\psi : \mathbf{R} \to \Theta$. Rather than using $\psi(w^T x)$ as the parameter value in our natural exponential family, we'll use $\psi(f(x))$, where f is some learned function. As before, lets take $\psi(\eta) = -e^\eta$. Then the likelihood for a single (x, y) pair is

$$e^{f(x)} e^{-\exp[f(x)]y}$$

   (where we're assuming that $y \geq 0$) and the loss (the negative log likelihood) is

$$-\left[f(x) - \exp\left[f(x)\right] y\right].$$

   Thus the empirical risk is (dropping the $1/n$ for convenience):

$$J(f) = \sum_{i=1}^{n} \left[\exp\left[f(x_i)\right] y_i - f(x_i)\right].$$

   For gradient boosting, we need to compute the gradient at the datapoints. So

$$\frac{\partial}{\partial f(x_i)} J(f) = \exp\left[f(x_i)\right] y_i - 1.$$

   So the algorithm is

   (a) Initialize $f_0(x) = 0$.

   (b) For $m = 1$ to M:

      i. Compute:

$$\mathbf{g}_m = \left(\exp\left[f_{m-1}(x_i)\right] y_i - 1\right)_{i=1}^n$$

      ii. Fit regression model to $-\mathbf{g}_m$:

$$h_m = \arg\min_{h \in \mathcal{F}} \sum_{i=1}^{n} \left((-\mathbf{g}_m)_i - h(x_i)\right)^2.$$

      A. Choose fixed step size $\nu_m = \nu \in (0, 1]$, or take

$$\nu_m = \arg\min_{\nu > 0} \sum_{i=1}^{n} \ell\left\{y_i, f_{m-1}(x_i) + \nu h_m(x_i)\right\}.$$

      B. Take the step:

$$f_m(x) = f_{m-1}(x) + \nu_m h_m(x)$$

8 Feature normalization when we regularize

9 Decision Trees

9.1 (10 min, 10 points) Building Trees by Hand[1]

In this problem we're going to be build a a small decision tree by hand for predicting whether or not a mushroom is poisonous. The training dataset is given below:

Poisonous	Size	Spots	Color
N	5	N	White
N	2	Y	White
N	2	N	Brown
N	3	Y	Brown
N	4	N	White
N	1	N	Brown
Y	5	Y	White
Y	4	Y	Brown
Y	4	Y	Brown
Y	1	Y	White
Y	1	Y	Brown

We're going to build a binary classification tree using the Gini index as the node impurity measure. The feature "Size" should be treated as numeric (i.e. we should find real-valued split points). For a given split, let R_1 and R_2 be the sets of data indices in each of the two regions of the split. Let $\hat{p}_1$ be the proportion of poisonous mushrooms in R_1, and let $\hat{p}_2$ be the proportion in R_2. Let N_1 and N_2 be the total number of training points in R_1 and R_2, respectively. Then the Gini index for the first region is $Q_1 = 2\hat{p}_1(1-\hat{p}_1)$ and $Q_2 = 2\hat{p}_2(1-\hat{p}_2)$ for the second region. When choosing our splitting variable and split point, we're looking to minimize the weighted impurity measure:

$$N_1 Q_1 + N_2 Q_2.$$

1. What is the first split for a binary classification tree on this data, using the Gini index? Work this out "by hand", and show your calculations. [Hint: This should only require calculating 6 weighted impurity measures.]

2. Compute the full decision tree by hand, building until all terminal nodes are either completely pure, or we cannot split any further.

3. Suppose we built the same type of tree described above (binary, Gini criterion, terminal nodes are either pure or cannot be split further) on the dataset given below. What would the training error be, given as a percentage? Why? [Hint: You can do this by inspection, without any significant calculations.]

[1] Based on Homework #4 from David Sontag's DS-GA 1003, Spring 2014.

Y	A	B	C
0	0	0	0
0	0	0	1
0	0	1	0
0	0	1	0
0	0	1	1
1	0	1	1
0	1	0	0
1	1	0	1
1	1	1	0
0	1	1	1
1	1	1	1

10 Kernels: short answer

1. show kernel as inner product is symmetric
2. kernel matrix is symmetric

A Convexity

A.1 Examples of Convex Functions (BV 3.1.5)

Functions mapping from $\mathbf{R}$:

- $x \mapsto e^{ax}$ is convex on $\mathbf{R}$ for all $a \in \mathbf{R}$
- $x \mapsto x^a$ is convex on $\mathbf{R}_{++}$ when $a \geq 1$ or $a \leq 0$ and concave for $0 \leq a \leq 1$
- $|x|^p$ for $p \geq 1$ is convex on $\mathbf{R}$
- $\log x$ is concave on $\mathbf{R}^{++}$
- $x \log x$ (either on $\mathbf{R}_{++}$ or on $\mathbf{R}_{+}$ if we define $0 \log 0 = 0$) is convex

Functions mapping from $\mathbf{R}^n$:

- Every norm on $\mathbf{R}^n$ is convex
- Max: $(x_1, \ldots, x_n) \mapsto \max\{x_1 \ldots, x_n\}$ is convex on $\mathbf{R}^n$
- Log-Sum-Exp[2]: $(x_1, \ldots, x_n) \mapsto \log(e^{x_1} + \cdots + e^{x_n})$ is convex on $\mathbf{R}^n$.

[2]This function can be interpreted as a differentiable (in fact, analytic) approximation to the max function, since

$$\max\{x_1, \ldots, x_n\} \leq \log(e^{x_1} + \cdots + e^{x_n}) \leq \max\{x_1, \ldots, x_n\} + \log n.$$

Can you prove it? Hint: $\max(a, b) \leq a + b \leq 2\max(a, b)$.

A.2 Operations the preserve convexity (BV 3.2, p. 79)

A.2.1 Nonnegative weighted sums

If $f_1, \ldots, f_m$ are convex and $w_1, \ldots, w_m \geq 0$, then $f = w_1 f_1 + \cdots + w_m f_m$ is convex. More generally, if $f(x, y)$ is convex in x for each $y \in \mathcal{A}$, and if $w(y) \geq 0$ for each $y \in \mathcal{A}$, then the function

$$g(x) = \int_{\mathcal{A}} w(y) f(x, y)\, dy$$

is convex in x (provided the integral exists).

A.2.2 Composition with an affine mapping

A function $f : \mathbf{R}^n \to \mathbf{R}^m$ is an **affine function** (or **affine mapping)** if it is a sum of a linear function and a constaint. That is, if it has the form $f(x) = Ax + b$, where $A \in \mathbf{R}^{m \times n}$ and $b \in \mathbf{R}^m$.

Composition of a convex function with an affine function is convex. More precisely: suppose $f : \mathbf{R}^n \to \mathbf{R}$, $A \in \mathbf{R}^{n \times m}$ and $b \in \mathbf{R}^n$. Define $g : \mathbf{R}^m \to \mathbf{R}$ by

$$g(x) = f\left(Ax + b\right),$$

with $\mathrm{dom} g = \{x \mid Ax + b \in \mathrm{dom} f\}$. Then if f is convex, then so is g; if f is concave, so is g. If f is **strictly** convex, and A has linearly independent columns, then g is also strictly convex.

A.2.3 Simple Composition Rules

- If g is convex then $\exp g(x)$ is convex.
- If g is convex and nonnegative and $p \geq 1$ then $g(x)^p$ is convex.
- If g is concave and positive then $\log g(x)$ is concave
- If g is concave and positive then $1/g(x)$ is convex.

A.2.4 Maximum of convex functions is convex (BV Section 3.2.3, p. 80)

Note: Below we use this to prove that the Lagrangian dual function is concave.

If $f_1, \ldots, f_m : \mathbf{R}^n \to \mathbf{R}$ are convex, then their pointwise maximum

$$f(x) = \max\left\{f_1(x), \ldots, f_m(x)\right\}$$

is also convex with domain $\mathrm{dom} f = \mathrm{dom} f_1 \cap \cdots \cap \mathrm{dom} f_m$.

This result extends to the supremum over arbitrary sets of functions (including uncountably infinite sets).

11-04-2015

References