STOR 565 Homework: Classification and Regression

1. Some general questions about rooted binary trees. Refer to the notes on clustering for the definitions.

   a. Draw a rooted binary tree with 3 nodes. How many leaves does it have? How many internal nodes does it have?

   b. Draw a rooted binary tree with 5 nodes. How many leaves does it have? How many internal nodes does it have?

   c. Draw essentially different rooted binary trees with 7 nodes. Do they have the same number of internal nodes? Do they have the same number of leaves?

   d. Formulate a conjecture about the relationship between the number of internal nodes and the number of leaves in a rooted binary tree.

   e. (Optional) Prove your conjecture using induction.

2. Let (X, Y) be a jointly distributed pair with $X \in \mathcal{X}$ and $Y \in \{0, 1\}$. Suppose that $\mathcal{X}$ is finite and that (X, Y) has joint probability mass function $p(x, y)$.

   a. Express the prior probabilities $\pi_0 = \mathbb{P}(Y = 0)$ and $\pi_1 = \mathbb{P}(Y = 1)$ in terms of $p(x, y)$.

   b. Express the class conditional probability mass function $p_0(x) = \mathbb{P}(X = x \,|\, Y = 0)$ in terms of $p(x, y)$ and the prior probabilities.

   c. Show that the marginal pmf of X can be written as $p(x) = \pi_0\, p_0(x) + \pi_1\, p_1(x)$ where $p_1(x) = \mathbb{P}(X = x \,|\, Y = 1)$.

   e. Use Bayes rule to show that $\eta(x) := P(Y = 1 \,|\, X = x) = \pi_1 p_1(x)/p(x)$

3. Consider a classification problem in which the predictor X is uniformly distributed on the unit interval $[0, 1]$ and the response $Y \in \{0, 1\}$ as usual. For $x \in [0, 1]$ let $\eta(x) = \mathbb{P}(Y = 1 \,|\, X = x)$. Specify the Bayes rule ϕ^* and the Bayes risk R^* in each of the following cases.

   a. $\eta(x) = 1/3$ for all x

   b. $\eta(x) = x$

c. $\eta(x) \in \{0, 1\}$ for all x

In each of the cases above, find the prior probability $\pi_1 = \mathbb{P}(Y = 1)$, or indicate why this is not possible without more information.

4. Let $(X, Y) \in \mathbb{R}^2 \times \{0, 1\}$ be a random predictor-response pair. Suppose that the predictor X is a pair (X_1, X_2) where $X_1, X_2 \in [0, 1]$ are independent, X_1 is uniform on $[0, 1]$, and X_2 has density $g(x_2) = 3x_2^2$ for $0 \leq x_2 \leq 1$. Suppose that $\eta(x_1, x_2) = (x_1 + x_2)/2$.

a. Find the Bayes rule ϕ^* for this problem and identify its decision boundary.

b. Find the unconditional density of X

c. Find the Bayes risk associated with (X, Y)

d. Find the prior probability that $Y = +1$.

e. Find the class-conditional density of X given $Y = 1$.

5. Consider the labeled data set $(-2, 1), (-1, 1), (0, 0), (1, 1), (2, 0) \in \mathbb{R} \times \{0, 1\}$.

a. Sketch the 1-nearest neighbor rule for this dataset by drawing a line and indicating which points are assigned to zero and which are assigned to one.

b. Sketch the 3-nearest neighbor rule for this dataset by drawing a line and indicating which points are assigned to zero and which are assigned to one.

6. Suppose that you are given access to a database consisting of many email messages that have been labeled as spam or normal. You decide to construct a simple classification rule, the only feature being whether or not the word "meeting" appears somewhere in the email. Using relative frequencies to estimate probabilities you find the following:

$$\hat{P}(\text{spam}) = .3 \quad \hat{P}(\text{`meeting' present} \,|\, \text{spam}) = .01 \quad \hat{P}(\text{`meeting' present} \,|\, \text{normal}) = .04$$

Using this information, calculate a simple classification rule for spam detection. What can you say about the error rate of your rule on the database?

*7. Argue as carefully as you can that if the Bayes risk R^* for a pair (X, Y) is equal to $1/2$ then Y is independent of X. Hint: Use the results of earlier HW problems.

8. Let $(X, Y) \in \mathbb{R} \times \{0, 1\}$ be a random predictor-response pair. Suppose that Y has prior probabilities $\pi_1 = \mathbb{P}(Y = 1)$ and $\pi_0 = \mathbb{P}(Y = 0)$, and that X is continuous with marginal density f and class conditional densities f_0 and f_1. Let $\eta(x) = \mathbb{P}(Y = 1 \,|\, X = x)$ as usual.

a. Show that the Bayes rule ϕ^* can be written in the form

$$\phi^*(x) \;=\; \begin{cases} 1 & \text{if } \log \frac{\eta(x)}{1-\eta(x)} \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

b. Find a simple expression for the Bayes rule $\phi^*(x)$ in terms of $\pi_1 f_1(x)$ and $\pi_0 f_0(x)$.

Suppose that f_1 is $\mathcal{N}(\mu_1, \sigma^2)$ and that f_0 is $\mathcal{N}(\mu_0, \sigma^2)$ where $\mu_1 > \mu_0$.

c. Using the results above, find an expression for the Bayes rule $\phi^*(x)$ in terms of the parameters π_0, π_1, μ_0, μ_1, and σ^2.

d. What is the form of the rule in part (b) when $\pi_1 = 1/2$? Explain why this makes intuitive sense.

e. Suppose for simplicity that $\mu_1 = u$ and $\mu_0 = -u$ for some $u > 0$. What form does the Bayes rule take when u increases (tends to infinity), and in particular, how does the rule depend on π_1 versus π_0? A informal but clear answer is fine.

9. Let $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R} \times \{0, 1\}$ be a labeled set of real observations.

a. Give an estimate of the probability that $Y = 0$. What does the law of large numbers say about the limiting behavior of your estimate as n gets very large?

b. Write the sample mean $\hat{\mu}_0$ of the zero-labeled observations using indicator functions.

c. Write the sample variance $\hat{\sigma}_0^2$ of the zero-labeled observations using $\hat{\mu}_0$ and indicator functions.

10. Consider the setting of linear discriminant analysis in which the class-conditional densities f_0 and f_1 have the multivariate normal form $f_k = \mathcal{N}(\mu_k, \Sigma_k)$.

a. Using the expression for the multivariate normal density, show that the discriminant functions $\delta_k(x) = \log(\pi_k \, f_k(x))$ have the form

$$\delta_k(x) = -\frac{1}{2} x^t \Sigma_k^{-1} x + \langle x, \Sigma_k^{-1} \mu_k \rangle - \frac{1}{2} \left[\log(2\pi)^d \pi_k^{-2} \det(\Sigma_k) + \mu_k^t \Sigma_k^{-1} \mu_k \right]$$

b. Show that when $\Sigma_0 = \Sigma_1 = \Sigma$ the decision boundary $B = \{x : \delta_1(x) = \delta_0(x)\}$ has the form

$$B \;=\; \{x : x^t\,\Sigma^{-1}(\mu_1 - \mu_0) + (c_0 - c_1) = 0\}$$

where c_0, c_1 are real valued constants, and argue that this set is a hyperplane.

11. Describe and discuss linear discriminant analysis.

12. Let (X, Y) be a jointly distributed pair with $X \in \mathbb{R}^d$ and $Y \in \{0, 1\}$. Suppose that we have added a zeroth component to the vector X that is always equal to 1, so that the augmented vector $X \in \mathbb{R}^{d+1}$. The logistic regression method for binary classification is based on the assumption that

$$\log \frac{\mathbb{P}(Y = 1 \,|\, X = x)}{\mathbb{P}(Y = 0 \,|\, X = x)} \;=\; \log \frac{\eta(x)}{1 - \eta(x)} \;=\; \langle \beta, x \rangle \tag{1}$$

for some vector $\beta \in \mathbb{R}^{d+1}$ of coefficients. In words, equation (1) says that the conditional log-odds ratio of $Y = 1$ vs. $Y = 0$ is linear in the feature vector x.

a. Show, by inverting the relation (1), that

$$\eta(x) = \eta(x : \beta) \;=\; \frac{e^{\langle \beta, x \rangle}}{1 + e^{\langle \beta, x \rangle}} \;=\; \frac{1}{1 + e^{-\langle \beta, x \rangle}}$$

Here we write $\eta(x : \beta)$ to remind ourselves that η depends on β.

b. Equation (1) is sometimes written in the form $\text{logit}(\eta(x)) = \langle \beta, x \rangle$, where $\text{logit}(u) = \log[u/(1-u)]$ for $0 < u < 1$ is the logistic (or logit) function. Sketch the logistic function.

Given a data set $D_n = (x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^{d+1} \times \{0, 1\}$ logistic regression estimates the coefficient vector β in (1) by maximizing the conditional log likelihood function

$$\ell(\beta) \;=\; \log \prod_{i=1}^{n} \mathbb{P}_\beta(Y = y_i \,|\, X = x_i)$$

where $\mathbb{P}_\beta(Y = 1 \,|\, X = x) = \eta(x : \beta)$ and $\mathbb{P}_\beta(Y = 0 \,|\, X = x) = 1 - \eta(x : \beta)$.

c. Use the expression for $\eta(x : \beta)$ in (a) to show that the conditional log likelihood function can be written in the form

$$\ell(\beta) \;=\; \sum_{i=1}^{n} \Big[y_i \langle \beta, x_i \rangle - \log(1 + e^{\langle \beta, x_i \rangle}) \Big]$$

d. Show that $\nabla \ell(\beta) = \sum_{i=1}^{n} x_i \, [y_i - \eta(x_i : \beta)]$. Hint: Evaluate the partial derivative $\partial \ell(\beta)/\partial \beta_j$ for a fixed index j between 1 and d.

13. Describe the difference between a fixed classification rule and a classification procedure. Define and discuss the conditional and expected risk of a classification procedure.

14. Let $D_n = (X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \{0, 1\}$ be iid observations for a classification problem. Recall that the empirical risk of a fixed classification rule $\phi : \mathcal{X} \to \{0, 1\}$ is defined by

$$\hat{R}_n(\phi) \;=\; \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\phi(X_i) \neq Y_i)$$

and that the risk of ϕ is $R(\phi) = \mathbb{P}(\phi(X) \neq Y)$.

a. Show that $\mathbb{E}[\hat{R}_n(\phi)] = R(\phi)$

b. Show that $\mathrm{Var}(\hat{R}_n(\phi)) = n^{-1} \, R(\phi)(1 - R(\phi)) \leq 1/(4n)$

c. Argue carefully that $n\hat{R}_n(\phi)$ has a $\mathrm{Bin}(n, R(\phi))$ distribution

d. Use Chebyshev's inequality to show that for $t \geq 0$

$$\mathbb{P}(|\hat{R}_n(\phi) - R(\phi)| \geq t) \;\leq\; \frac{R(\phi)(1 - R(\phi))}{n \, t^2} \;\leq\; \frac{1}{4 \, n \, t^2}$$

e. Use Hoeffding's inequality to show that for $t \geq 0$

$$\mathbb{P}(|\hat{R}_n(\phi) - R(\phi)| \geq t) \;\leq\; 2 \exp\{-2nt^2\}$$

15. Consider a classification problem in which you have access to a test set containing $m = 120$ iid observations $(X_i, Y_i) \in \mathcal{X} \times \{0, 1\}$. You would like to use the test set to assess the risk of a given rule ϕ using the empirical risk $\hat{R}_m(\phi)$. Chebyshev's inequality and Hoeffding's inequality provide bounds on $\mathbb{P}(|\hat{R}_m(\phi) - R(\phi)| \geq t)$ for $t \geq 0$. Compute and compare these probability bounds, with $m = 120$, at the following values of t: 1/20, 1/11, 1/9, and 1/5.

16. Consider a classification problem in which you would like to assess the risk of a given rule ϕ using its empirical risk $\hat{R}_m(\phi)$ on a test data set D_m. In particular, you wish to determine the size n of the test set necessary to conclude that

$$\mathbb{P}(|\hat{R}_n(\phi) - R(\phi)| \geq \delta) \leq \epsilon$$

Use Chebyshev's and Hoeffding's inequalities to find suitable values for n as a function of δ and ϵ. How do the resulting quantities depend on δ and ϵ? Generally speaking, which inequality permits you to use a smaller test set?

17. Let D_n and D_m be independent training and test sets, respectively. Suppose that the rule $\hat{\phi}_n(x) = \phi_n(x : D_n)$ is derived from the training set.

a. Define the test set error $\hat{R}_m(\hat{\phi}_n)$.

b. Show that $\mathbb{E}[\hat{R}_m(\hat{\phi}_n) \mid D_n] = R(\hat{\phi}_n)$

c. What is $\mathbb{E}\hat{R}_m(\hat{\phi}_n)$? Compare this to your answer above.

18. Let $(x, y) \in \mathbb{R}^p \times \mathbb{R}$ be a fixed predictor-response pair, and define a function $f : \mathbb{R}^p \to \mathbb{R}$ by $f(\beta) = (y - x^t\beta)^2$.

a. Show that f is convex.

b. Now let $D_n = (x_1, y_1), \ldots, (x_n, y_n)$ be n predictor-response pairs. What can you say about the convexity of the sum of squares $g(\beta) = \sum_{i=1}^n (y_i - x_i^t\beta)^2$?

c. Fix $\lambda \geq 0$ and define the penalized performance criterion

$$h_\alpha(\beta) \;=\; \sum_{i=1}^{n}(y_i - x_i^t\beta)^2 \;+\; \lambda \sum_{j=1}^{p} |\beta_j|^\alpha$$

Argue that h_α is convex if $\alpha \geq 1$. Hint: Recall that a sum of convex functions is convex.

19. Consider a data set with design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ and response vector $\mathbf{y} \in \mathbb{R}^n$. Fix $\lambda > 0$ and define the penalized loss $\hat{R}_{n,\lambda}(\beta) = ||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda\,||\beta||^2$. Following the calculus based arguments for OLS, show that $\hat{R}_{n,\lambda}(\beta)$ has unique minimizer $\hat{\beta}_\lambda = (\mathbf{X}^t\mathbf{X} + \lambda I_p)^{-1}\mathbf{X}^t y$.

20. Let $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$ be a jointly distributed pair following the signal plus noise model $Y = f(X) + \varepsilon$ where ε is independent of X, $\mathbb{E}\varepsilon = 0$, and $\mathrm{Var}(\varepsilon) = \sigma^2$.

a. Find simple expressions for $\mathbb{E}Y$ and $\mathrm{Var}(Y)$.

b. Argue that $\mathbb{E}(Y|X) = f(X)$. Thus f is the regression function of Y based on X.

c. Show that $\varphi = f$ minimizes the risk $R(\varphi) = \mathbb{E}(\varphi(X) - Y)^2$ over prediction rules $\varphi : \mathbb{R}^p \to \mathbb{R}$. What is the minimum value of $R(\varphi)$?

21. Let $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathbb{R}$ be iid observations from the signal plus noise model $Y = f(X) + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

a. Define the empirical risk $\hat{R}_n(\varphi)$ of a rule $\varphi : \mathbb{R}^p \to \mathbb{R}$.

b. Assuming that $\mathrm{Var}(\varphi(X)) < \infty$, find the expectation and variance of $\hat{R}_n(\varphi)$. You may use the fact that $\mathbb{E}\varepsilon^3 = 0$ and $\mathbb{E}\varepsilon^4 = 3\sigma^4$ under our normality assumption.

c. What does Chebyshev's inequality tell you in this setting? What sort of assumptions could you make to control the size of the upper bound?

d. Can you apply Hoeffding's inequality in this case? If so, what is the bound?

22. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^{p+1}$ be fixed vectors with initial component equal to one 1. Suppose that we observe responses $y_1, \ldots, y_n \in \mathbb{R}$ generated from the linear model $y_i = \beta^t \mathbf{x}_i + \varepsilon_i$, where $\beta \in \mathbb{R}^{p+1}$ is an unknown coefficient vector and $\varepsilon_1, \ldots, \varepsilon_n$ are iid $\sim \mathcal{N}(0, \sigma^2)$.

a. Argue that $y_1, \ldots, y_n$ are independent and that $y_i \sim \mathcal{N}(\mathbf{x}_i^t \beta, \sigma^2)$.

b. Find the joint likelihood $L(\beta)$ of $y_1, \ldots, y_n$.

c. Find the log likelihood $\ell(\beta)$ of $y_1, \ldots, y_n$ and show that maximizing $\ell(\beta)$ over β is equivalent to minimizing the empirical risk $\hat{R}_n(\beta) = n^{-1} \sum_{i=1}^n (y_i - \mathbf{x}_i^t \beta)^2$ over β.

d. Define the response vector $\mathbf{y}$ and design matrix $\mathbf{X}$ associated with the data above, giving the dimensions of each. Show carefully that $\hat{R}_n(\beta) = n^{-1} ||y - X\beta||^2$.

23. Let $\mathbf{y}$ and $\mathbf{X}$ be the response vector and design matrix, respectively, associated with observations $(\mathbf{x}_i, y_i)$ of the previous problem. Recall from class that the OLS coefficient $\hat{\beta} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{y}$

a. Show that $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$. Conclude that $\mathbf{y} \sim \mathcal{N}_n(\mathbf{X}\beta, \sigma^2 I)$.

b. Show that $\hat{\beta} = \beta + (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \varepsilon$.

c. Find $\mathbb{E}\hat{\beta}$ and $\mathrm{Var}(\hat{\beta})$.

d. Argue that $\hat{\beta} \sim \mathcal{N}_{p+1}(\beta, \sigma^2(\mathbf{X}^t\mathbf{X})^{-1})$, and conclude that $\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}_{jj})$.

e. Use the distribution of $\hat{\beta}_j$ to find a 95% confidence interval for β_j.

24. Let $\mathbf{y}$ and $\mathbf{X}$ be the response vector and design matrix, respectively, associated with observations $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n) \in \mathbb{R}^p \times \mathbb{R}$.

a. Show that $\mathbf{X}^t\mathbf{X} + \lambda I_p$ is symmetric and positive definite if $\lambda > 0$. Conclude that $\mathbf{X}^t\mathbf{X} + \lambda I_p$ is invertible if $\lambda > 0$.

b. Find a simple relationship between the eigenvalues of $\mathbf{X}^t\mathbf{X} + \lambda I_p$ and those of $\mathbf{X}^t\mathbf{X}$.

25. Let $\hat{\beta}_\lambda$ be the minimizer of $\hat{R}_{n,\lambda}(\beta) = ||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda\,||\beta||^2$.

a. Show that $\hat{\beta}_0$ is the usual OLS estimator (when the rank of X is equal to p).

b. Show that $||\mathbf{y} - \mathbf{X}\hat{\beta}_\lambda||^2 \leq ||\mathbf{y} - \mathbf{X}\beta||^2$ for every β such that $||\beta|| \leq ||\hat{\beta}_\lambda||$. Hint: Assume the stated inequality fails to hold and show that this implies that $\hat{\beta}_\lambda$ is not the minimizer of $\hat{R}_{n,\lambda}(\beta)$.

26. Let $D_n = (x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^p \times \{\pm 1\}$ be sequence of labeled pairs. Show that the constraint set

$$C := \{w, b : y_i(x_i^t w - b) \geq 1 \text{ for } i = 1, \ldots, n\}$$

appearing in the primal SVM optimization problem is convex. To make things a bit more formal, treat the elements of C as vectors $v = (w_1, \ldots, w_p, b)^t \in \mathbb{R}^{p+1}$. Hint: Show that C is the intersection of n sets, one for each i, and then show that each of these sets is convex.

In the next two questions you are asked to fill in some of the details from the SVM lecture concerning how one finds the maximum margin classifier for linearly separable data.

27. Write down the primal problem, with optimal value $p*$, and argue using the previous question and results from a previous homework that the primal problem is a convex program. Now consider the Lagrangian $L : \mathbb{R}^p \times \mathbb{R} \times \mathbb{R}^n_+$, which is defined by

$$L(w, b, \lambda) \;:=\; \frac{1}{2}||w||^2 \,-\, \sum_{i=1}^{n} \lambda_i \left\{y_i(w^t x_i - b) - 1\right\}$$

Establish that

$$\max_{\lambda \geq 0} L(w, b, \lambda) \;=\; \begin{cases} ||w||^2 & \text{if } y_i(x_i^t w - b) \geq 1 \text{ for } i = 1, \ldots, n \\ +\infty & \text{otherwise} \end{cases}$$

To see why this is true, note that if one of the constraints $y_i(x_i^t w - b) \geq 1$ is *not* satisfied, then one can increase the corresponding λ_i to make the Lagrangian arbitrarily large. Using the last display above, argue informally that the primal problem can be written as

$$p^* \;=\; \min_{w,b} \max_{\lambda \geq 0} L(w, b, \lambda)$$

28. Let $D_n = (x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \{0, 1\}$ be a data set for classification. For each region $A \subseteq \mathcal{X}$ let $|A|$ denote the number of points x_i in A and let $p(A) = |A|^{-1} \sum_{x_i \in A} y_i$ be the fraction of points $x_i \in A$ labeled 1. Suppose that the region A can be expressed as the disjoint union $A = A_1 \cup A_2$ of two other regions.

a. Using the definition, show that

$$p(A) \;=\; \frac{|A_1|}{|A|} p(A_1) + \frac{|A_2|}{|A|} p(A_2)$$

b. Show that $|A| = |A_1| + |A_2|$. Conclude from this and part (a) that for any concave function $f : [0, 1] \to \mathbb{R}$

$$f(p(A)) - \frac{|A_1|}{|A|} f(p(A_1)) - \frac{|A_2|}{|A|} f(p(A_2)) \;\geq\; 0$$

This establishes that the impurity differences defined in the lecture for the misclassification, Gini, and entropy impurity measures are non-negative.

c. Let $m(p) = \min(p, 1 - p)$. Show that $|A| m(p(A))$ is the number of misclassifications if every point in A is assigned to the majority class.

d. Consider two partitions γ_1 and γ_2 of $\mathcal{X}$ that are identical except that a cell A of γ_1 is split into two cells A_1 and A_2 in γ_2. What can you say about the training error of the corresponding histogram classification rules (based on majority voting in cells)?

29. Let $D_n = (x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \{0, 1\}$ be a data set for classification and let $\gamma = \{A_1, \ldots, A_K\}$ be a partition of $\mathcal{X}$. Define the histogram classification rule $\hat{\phi}_\gamma$ based on γ. Show that $\hat{\phi}_\gamma$ minimizes the training error $R_n(\phi)$ over all classification rules ϕ that are constant on the cells of γ, meaning $\phi(u) = \phi(v)$ if u, v are in the same cell of γ.

30. Recall that the Bayes Rule ϕ^* for a jointly distributed pair (X, Y) with response $Y \in \{0, 1\}$ is defined by

$$\phi^*(x) = \underset{k=0,1}{\operatorname{argmax}} \, \mathbb{P}(Y = k \,|\, X = x)$$

a. How would you modify this definition in the case where the response takes values in the finite set $\{0, 1, \ldots, K\}$, that is, each feature vector x is associated with one of K possible outcomes?

b. Show that in the binary case $Y \in \{0, 1\}$ the Bayes Rule has the form

$$\phi^*(x) \;=\; \begin{cases} 1 & \text{if } \eta(x) \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$$