

CS 229 Midterm Review Handout

1 Study list

You should study the following topics in preparation for the midterm:

- Linear regression
 - Stochastic versus batch gradient descent
 - Normal equations
 - Maximum likelihood interpretation
- Locally weighted linear regression
 - Understand that the weights are recalculated at each query point
- Logistic regression
 - Gradient descent versus Newton's method
 - How to show that the log likelihood is convex via the Hessian
- Exponential family
 - How to show a distribution is a member of the family
 - How to construct a GLM
- GDA
 - How to estimate
 - How to find the decision boundary
 - Comparison to logistic regression
- Naive Bayes
 - Multivariate Bernoulli event model
 - Multinomial event model
 - Laplace smoothing
- SVMs
 - Large margin
 - Lagrange duality, KKT conditions
 - Specifically, how KKT conditions result in support vectors
 - Regularized SVMs

- The SMO algorithm
- Kernels
 - Why the kernel trick is justified
 - How to kernelize an algorithm
 - Mercer’s theorem
- Learning theory
 - The bias/variance tradeoff
 - Union bound
 - Hoeffding inequality / Chernoff bound
 - Uniform convergence
 - Bounds on generalization
 - VC dimension
- MAP estimation and Bayesian statistics
- Debugging machine learning applications

2 Example problems

1. SVMs with no intercept term

In this problem we will consider Support Vector Machines where the classifier is of the form $f(x) = w^T x$ (i.e., without the $+b$ intercept term that we previously saw for regular SVMs). This leads to the following optimization problem

$$\begin{aligned} \min_{w, \xi} \quad & \frac{1}{2} w^T w + C \sum_{i=1}^m \xi_i \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)}) \geq 1 - \xi_i, \quad i = 1, \dots, m \\ & \xi_i \geq 0, \quad i = 1, \dots, m \end{aligned}$$

- (a) For logistic regression, we did not require an explicit intercept term, because we could define x_0 to equal 1, and let θ_0 play the role of the intercept. For SVMs, can we similarly define x_0 or $\phi_0(x)$ to be equal to 1, and use w_0 as the intercept term, or will this change the decision function found by the SVM?

Answer: This will change the decision function found by the SVM. In the original SVM formulation, the $\frac{1}{2}\|w\|^2$ term did not penalize the intercept b , so b would be chosen as large as possible to satisfy the constraints. When we treat w_0 as the intercept it will be regularized as well, changing the optimal solution.

- (b) We will now find the dual of the optimization problem above. First, write down the Lagrangian. Use α_i and s_i to denote the Lagrange multipliers corresponding to the first and second sets of constraints respectively in the primal optimization problem above.

Answer:

Recall from section last week that in Lagrange duality problems, we want the optimization problem to be of the form:

$$\begin{aligned} \min_x f(x) \\ \text{subject to } g_i(x) \leq 0, i = 1, \dots, m, \\ h_i(x) = 0, i = 1, \dots, p. \end{aligned}$$

Given an optimization problem of that form, the Lagrangian is defined as

$$\mathcal{L}(x, \alpha, \beta) = f(x) + \sum_{i=1}^m \alpha_i g_i(x) + \sum_{i=1}^p \beta_i h_i(x).$$

Our SVM problem needs to be rewritten slightly to be in standard form:

$$\begin{aligned} \min_{w, \xi} \quad & \frac{1}{2} w^T w + C \sum_{i=1}^m \xi_i \\ \text{s.t.} \quad & -y^{(i)}(w^T x^{(i)}) + 1 - \xi_i \leq 0, \quad i = 1, \dots, m \\ & -\xi_i \leq 0, \quad i = 1, \dots, m \end{aligned}$$

Now that we have the optimization problem in the correct form, we simply plug in the objective and the constraints into the formula on the Lagrangian:

$$\mathcal{L}(w, \xi, \alpha, s) = \frac{1}{2} w^T w + C \sum_{i=1}^m \xi_i + \sum_{i=1}^m \alpha_i (-y^{(i)}(w^T x^{(i)}) + 1 - \xi_i) + \sum_{i=1}^m s_i (-\xi_i)$$

- (c) In order to find the dual problem, calculate the following derivatives with respect to the primal variables

Answer:

$$\begin{aligned} (\nabla_w \mathcal{L}(w, \xi, \alpha, s))_j &= \frac{\partial \mathcal{L}}{\partial w_j} \left(\frac{1}{2} w^T w - \sum_{i=1}^m \alpha_i (y^{(i)}(w^T x^{(i)})) \right) \\ &= w_j - \sum_{i=1}^m \alpha_i y^{(i)} x_j^{(i)} \end{aligned}$$

$$\begin{aligned}
(\nabla_{\xi} \mathcal{L}(w, \xi, \alpha, s))_j &= \frac{\partial \mathcal{L}}{\partial \xi_j} (C \sum_{i=1}^m \xi_i - \sum_{i=1}^m \alpha_i \xi_i - \sum_{i=1}^m s_i \xi_i) \\
&= \frac{\partial \mathcal{L}}{\partial \xi_j} (C \cdot \mathbf{1}^T \xi - \alpha^T \xi - s^T \xi) \\
&= C - \alpha_j - s_j
\end{aligned}$$

- (d) Find the dual optimization problem. You should write down the dual optimization problem in the following form:

$$\max_{\alpha} W(\alpha) =$$

s.t.

Be sure to simplify your answer as much as possible (hint: use the KKT conditions). If simplified fully, your answer should not include any of the s_i Lagrange multipliers.

Answer: Recall from section last week that the general form of the Lagrange dual problem is

$$\max_{\alpha, \beta: \alpha_i \geq 0, \forall i} \theta_D(\alpha, \beta)$$

where

$$\theta_D(\alpha, \beta) = \min_x \mathcal{L}(x, \alpha, \beta)$$

To find θ_D , we need to minimize the Lagrangian with respect to w and ξ . In order to do so, we set the above derivatives to zero, resulting in

$$\begin{aligned}
w &= \sum_{i=1}^m \alpha_i y^{(i)} x^{(i)} \\
s_i &= C - \alpha_i
\end{aligned}$$

Substituting these back into the Lagrangian we get

$$\begin{aligned}\mathcal{L}(\alpha, s) &= \frac{1}{2} \left(\sum_{i=1}^m \alpha_i y^{(i)} x^{(i)} \right)^T \left(\sum_{i=1}^m \alpha_i y^{(i)} x^{(i)} \right) + C \sum_{i=1}^m \xi_i \\ &\quad - \sum_{i=1}^m \alpha_i \left(y^{(i)} \left(\sum_{j=1}^m \alpha_j y^{(j)} x^{(j)} \right)^T x^{(i)} - 1 + \xi_i \right) - \sum_{i=1}^m (C - \alpha_i) \xi_i \\ &= \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)}.\end{aligned}$$

It is worth working through the preceding step on your own because this sort of manipulation occurs frequently in this type of problem. It should look familiar—the same type of simplification step was necessary for the original SVM too.

We now have θ_D . The constraints for the dual problem are that all of the dual variables are non-negative:

$$\begin{aligned}0 &\leq \alpha_i, i = 1, \dots, m \\ 0 &\leq s_i, i = 1, \dots, m \\ s_i &= C - \alpha_i, i = 1, \dots, m\end{aligned}$$

At this point we have found the dual problem. However, we can still simplify it further by removing the s variables. Because we already found that $s_i = C - \alpha_i$ we can rewrite the constraints as:

$$0 \leq \alpha_i \leq C, i = 1, \dots, m$$

Writing out the optimization problem in its complete form, we have:

$$\begin{aligned}\max_{\alpha} \quad & W(\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)} \\ & 0 \leq \alpha_i \leq C, \quad i = 1, \dots, m\end{aligned}$$

- (e) Suppose we wanted to map each $x^{(i)}$ into a high-dimensional, possibly infinite-dimensional, feature space. Show how you would modify the optimization problem to keep training feasible, and how you would make a prediction on a new input x .

Answer: To modify the optimization problem, simply replace the inner product at the end with a kernel.

To classify a new input x , we want to find the sign of $w^T \phi(x)$, but both of these quantities are potentially infinite dimensional. We need to find a way to express $w^T x$ using only inner products over feature vectors. From our answer to part c, we know that

$$\nabla_w \mathcal{L}(w, \xi, \alpha, s) = w - \sum_{i=1}^m \alpha_i y^{(i)} x^{(i)}.$$

Since the gradient must be zero, this implies that

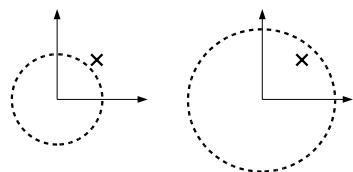
$$w^T x = \sum_{i=1}^m \alpha_i y^{(i)} x^{(i)T} x = \sum_{i=1}^m \alpha_i y^{(i)} K(x^{(i)}, x)$$

2. V.C. dimension

Let the input domain of a learning problem be $\mathcal{X} = \mathbb{R}^2$. Give the VC dimension for each of the following classes of hypotheses. Your proofs do not need to be completely rigorous, but if you claim, for example, the VC dimension of some hypothesis class is d , then give a brief explanation of why that hypothesis class can shatter some d points, but why there are no $d + 1$ points that it can shatter.

- Circles centered at the origin: $h(x) = 1\{\|x\| < r\}$ with parameter $r \in \mathbb{R}$.

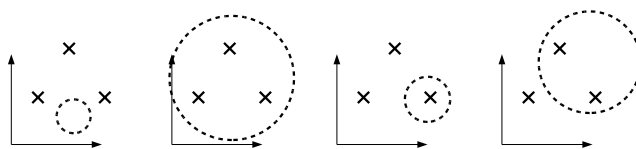
Answer: VC dimension = 1.



(Note that circles centered at the origin cannot shatter two points, since we could never label the closer point negative and the farther point positive)

- Circles with arbitrary origin: $h(x) = 1\{\|x - x_c\| < r\}$ with parameters $r \in \mathbb{R}$, $x_c \in \mathbb{R}^2$.

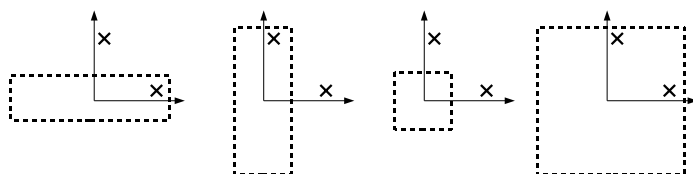
Answer: VC dimension = 3.



(And similar hypotheses).

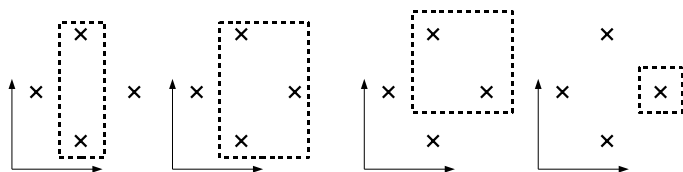
- Rectangles centered at the origin: $h(x) = 1\{-a < x_1 < a, -b < x_2 < b\}$ with parameters $a, b \in \mathbb{R}$.

Answer: VC dimension = 2.



- Rectangles with arbitrary origin: $h(x) = 1\{a < x_1 < b, c < x_2 < d\}$ with parameters $a, b, c, d \in \mathbb{R}$.

Answer: VC dimension = 4.



(And other similar hypotheses).

3. Learning Theory

In class we derived error bounds for the Empirical Risk Minimization (ERM) algorithm, which picks the hypothesis $h \in \mathcal{H}$ that minimizes training error. However, finding such a hypothesis can be difficult, so in this problem we will consider an *approximate* Empirical Risk Minimization algorithm, which can only find a hypothesis that attains close to the minimum possible training error. Formally, for some fixed parameters $\sigma > 1$ and $\rho \in [0, 1]$, approximate ERM will return a hypothesis $\hat{h} \in \mathcal{H}$ that satisfies

$$\hat{\varepsilon}(\hat{h}) \leq \sigma \cdot \left(\min_{h \in \mathcal{H}} \hat{\varepsilon}(h) \right) + \rho.$$

Derive a bound on the *generalization* error of the approximate ERM algorithm. For this question you may assume that \mathcal{H} is finite, with $|\mathcal{H}| = k$. Your bound should state that with probability at least $1 - \delta$,

$$\varepsilon(\hat{h}) \leq f(k, m, \delta, \varepsilon(h^*), \sigma, \rho)$$

for some function f , where, h^* denotes the hypothesis with minimum generalization error. When $\sigma = 1$ and $\rho = 0$, this should reduce to the bound given in class.

Answer: By applying the Hoeffding inequality, we have that for

$$\gamma = \sqrt{\frac{1}{2m} \log \frac{2k}{\delta}}$$

with probability at least $1 - \delta$,

$$\begin{aligned} \varepsilon(\hat{h}) &\leq \hat{\varepsilon}(\hat{h}) + \gamma \\ &\leq \sigma \hat{\varepsilon}(h^*) + \rho + \gamma \\ &\leq \sigma(\varepsilon(h^*) + \gamma) + \rho + \gamma. \end{aligned}$$

Therefore the final bound is that, with probability $1 - \delta$,

$$\varepsilon(\hat{h}) \leq \sigma \varepsilon(h^*) + (1 + \sigma) \sqrt{\frac{1}{2m} \log \frac{2k}{\delta}} + \rho.$$