CS 189: Introduction to Machine Learning - Discussion 2

1. Fun with the SVM margin

   a. We typically frame an SVM problem as trying to *maximize* the margin. Explain intuitively why a bigger margin will result in a model that will generalize better, or perform better in practice.

   Solution:

   1. One intuition is that if points are closer to the border, we are less certain about their class. Thus, it would make sense to create a boundary where our "certainty" is highest about all the training set points.
   2. Another intuition involves thinking about the process that generated the data we are working with. Since it's a noisy process, if we drew a boundary close to one of our training points of some class, it's very possible that a point of the same class will be generated across the boundary, resulting in an incorrect classification. Therefore it makes sense to make the boundary as far away from our training points as possible.

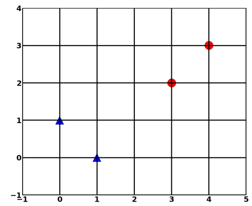
   b. Show that the width of an SVM margin with linearly separable data is $\frac{2}{\|w\|}$.

   Solution: The width of the margin is defined by the points that lie on it, also called support vectors. Let's say we have a point, $\vec{x}'$, which is a support vector. The distance between $\vec{x}'$ and the separating hyperplane can be calculated by projecting the vector starting at the plane and ending at $\vec{x}$ onto the plane's unit normal vector. The equation of the plane is $\vec{w}^T\vec{x} + b = 0$. Since $\vec{w}$ by definition is orthogonal to the hyperplane, we want to project $\vec{x}' - \vec{x}$ onto the unit vector normal to the hyperplane, $\frac{\vec{w}}{\|\vec{w}\|}$.

   $$\frac{\vec{w}^T}{\|\vec{w}\|}(\vec{x}' - \vec{x}) = \frac{1}{\|\vec{w}\|}(\vec{w}^T\vec{x}' - \vec{w}^T\vec{x}) = \frac{1}{\|\vec{w}\|}(\vec{w}^T\vec{x}' + b - \vec{w}^T\vec{x} - b)$$

   Since we set $\vec{w}^T\vec{x}' + b = 1$ (or -1) and by definition, $\vec{w}^T\vec{x} + b = 0$, this quantity just turns into $\frac{1}{\|\vec{w}\|}$, or $\vec{1}\|\vec{w}\|$, so the distance is the absolute value, $\frac{1}{\|\vec{w}\|}$. Since this is half the margin, we double it to get the full width of $\frac{2}{\|\vec{w}\|}$.

   c. You're presented with the following set of data (triangle = +1, circle = -1):



Find the equation (by hand) of the hyperplane $\vec{w}^T x + b = 0$ that would be used by an SVM classifier. Which points are support vectors?

Solution: The equation of the hyperplane will pass through point $(2, 1)$, with a slope of -1, since it's the halfway point between the two closest points between the classes. The equation of this line is $x_1 + x_2 = 3$. We know that from this form, $w_1 = w_2$. We also know that the at the support vectors, $w^T x + b = \pm 1$. This gives us the equations:

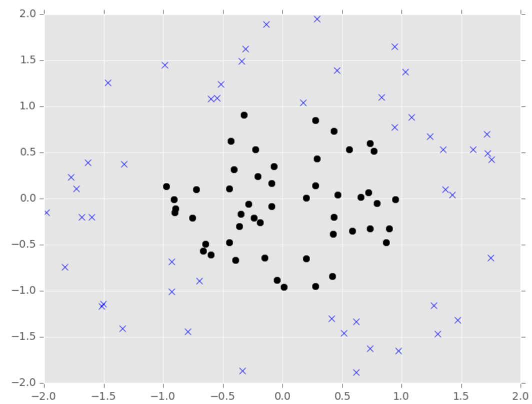
$$1w_1 + 0w_2 + b = 1$$

$$3w_1 + 2w_2 + b = -1$$

Solving this system of equations, we get $\vec{w} = [-\frac{1}{2}, -\frac{1}{2}]^T$ and $b = \frac{3}{2}$.
The support vectors are $(1, 0), (0, 1)$, and $(3, 2)$.

2. Circular Distributions

Consider the following dataset where each point $x_n = (x_{1,n}, x_{2,n})$ is sampled *iid* and



uniformly at random from two equiprobable (each equally likely) classes, a disk of radius 1 ($y_n = 1$) and a ring from 1 to 2 ($y_n = -1$).

Qualitatively estimate and sketch the following quantities

- The class conditional density $P(X \mid Y)$
- The density of X
- The conditional density of $P(Y \mid X)$
- The Bayes Classifier
- The Bayes Risk

Solution:

- $$f_{X|Y=1} = \begin{cases} \frac{1}{\pi} & \|X\|_2 \leq 1 \\ 0 & o.w. \end{cases}$$

  $$f_{X|Y=-1} = \begin{cases} \frac{1}{3\pi} & 1 \leq \|X\|_2 \leq 2 \\ 0 & o.w. \end{cases}$$

- $$f_X = \frac{1}{2}\left(f_{X|Y=1} + f_{X|Y=-1}\right)$$

- $$\mathbb{P}\left[Y=1 \mid X\right] = \begin{cases} 1 & \|X\|_2 \le 1 \\ 0 & o.w. \end{cases}$$

  $$\mathbb{P}\left[Y=-1 \mid X\right] = \begin{cases} 1 & 1 \le \|X\|_2 \le 2 \\ 0 & o.w. \end{cases}$$

- Bayes Classifier: the circle $x_1^2 + x_2^2 = 1$
- The Bayes Risk $= 0$

Now suppose you're given only one of the axes, namely $x_{1,n}$, re-estimate these quantities.

Solution:

- $$f_{X|Y=1} = \begin{cases} \frac{2}{\pi}\sqrt{1-x_1^2} & \|x_1\|_2 \le 1 \\ 0 & o.w. \end{cases}$$

  $$f_{X|Y=-1} = \begin{cases} \frac{2}{3\pi}\sqrt{4-x_1^2} & 1 \le \|x_1\|_2 \le 2 \\ \frac{2}{3\pi}\left(\sqrt{4-x_1^2} - \sqrt{1-x_1^2}\right) & \|x_1\|_2 \le 1 \\ 0 & o.w. \end{cases}$$

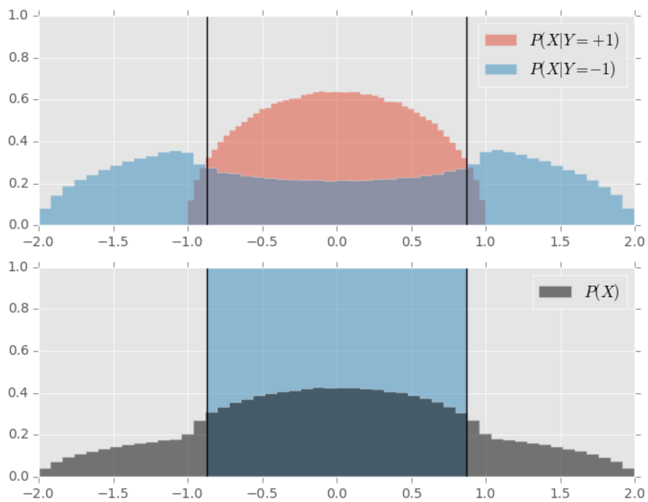
- $$\begin{aligned} f_X &= \frac{1}{2}\left(f_{X|Y=1} + f_{X|Y=-1}\right) \\ &= \begin{cases} \frac{1}{3\pi}\sqrt{4-x_1^2} & 1 \le \|x_1\|_2 \le 2 \\ \frac{1}{3\pi}\left(\sqrt{4-x_1^2} + 2\sqrt{1-x_1^2}\right) & \|x_1\|_2 \le 1 \\ 0 & o.w. \end{cases} \end{aligned}$$

- $$\mathbb{P}\left[Y=1 \mid X\right] = \begin{cases} 1 - \frac{\sqrt{4-x_1^2}-\sqrt{1-x_1^2}}{\sqrt{4-x_1^2}+2\sqrt{1-x_1^2}} & \|x_1\|_2 \le 1 \\ 0 & o.w. \end{cases}$$

  $$\mathbb{P}\left[Y=-1 \mid X\right] = \begin{cases} 1 & 1 \le \|x_1\|_2 \le 2 \\ \frac{\sqrt{4-x_1^2}-\sqrt{1-x_1^2}}{\sqrt{4-x_1^2}+2\sqrt{1-x_1^2}} & \|x_1\|_2 \le 1 \\ 0 & o.w. \end{cases}$$

- Bayes Classifier: $\hat{y} = 1$ if $\|x_1\|_2 \leq \sqrt{\frac{4}{5}} = 0.894$
- The Bayes Risk

$$R^* = 2\left(\int_0^{\sqrt{4/5}} \mathbb{P}\left[Y = -1 \mid X = x\right] f_X(x)dx + \int_{\sqrt{4/5}}^{1} \mathbb{P}\left[Y = 1 \mid X = x\right] f_X(x)dx\right)$$
$$= 2(0.10 + 0.01) = 0.22$$



3. Bayesian Decision Theory: Case Study - We're going fishing!

   We want to design an automated fishing system that captures fish, classifies them, and sends them off to two different companies, Salmonites, Inc., and Seabass, Inc. For some reason we only ever catch salmon and seabass. Salmonites, Inc. wants salmon, and Seabass, Inc. wants seabass. Given only the weights of the fish we catch, we want to figure out what type of fish it is using machine learning!

   Let us assume that the weight of both seabass and salmon are both normally distributed (univariate Gaussian), given by the p.d.f.:

$$P(x|\mu_i, \sigma_i) = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}$$

   We are given this data:

Data for salmon: $\{3, 4, 5, 6, 7\}$

Data for seabass: $\{5, 6, 7, 8, 9, 7+\sqrt{2}, 7-\sqrt{2}\}$

When we classify seabass incorrectly, it gets sent to Salmonites, Inc. who won't pay us for the wrong fish and sells it themselves. When we classify salmon incorrectly, it gets sent to SeaBass, Inc., who is nice and returns our fish. This situation gives rise to this loss matrix:

Predicted:

Truth:		**salmon**	**seabass**
	salmon	0	1
	seabass	2	0

a) First, compute the ML estimates for μ and σ for the univariate Gaussian in both the seabass and the salmon case. Also compute the empirical estimates of the priors. (Salmon = 1, Seabass = 2)

Solution:

$$L(X_1, X_2, ..., X_N; \mu, \sigma) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N e^{-\frac{1}{2\sigma^2}\sum\limits_i (X_i-\mu)^2}$$

$$l(X_1, X_2, ..., X_N; \mu, \sigma) = N\ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{1}{2\sigma^2}\sum_i (X_i-\mu)^2$$

Solving for μ.

$$\frac{\partial l(\mu,\sigma)}{\partial \mu} = \frac{1}{\sigma^2}\left(\sum_i X_i - N\mu\right) = 0$$

$$\hat{\mu} = \frac{1}{N}\sum_i X_i$$

Solving for σ.

$$\frac{\partial l(\mu,\sigma)}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_i (X_i-\mu)^2 = 0$$

$$\hat{\sigma} = \sqrt{\frac{1}{N}\sum_i (X_i-\hat{\mu})^2}$$

Plugging in numbers for seabass and salmon: $\mu_1 = 5$, $\mu_2 = 7$, $\sigma_1 = \sqrt{2}$, $\sigma_2 = \sqrt{2}$

Calculating the priors: $\pi_1 = 5/12$, $\pi_2 = 7/12$

$\hat{\mu_1} =$ $\hat{\mu_2} =$
$\hat{\sigma_1} =$ $\hat{\sigma_2} =$
$\hat{\pi_1} =$ $\hat{\pi_2} =$

What is significant about $\hat{\sigma_1}$ and $\hat{\sigma_2}$?

Solution: They're the exact same, so a decision boundary between the two Gaussians characterized by them will be linear.

b) Next, find the decision rule when assuming a 0-1 loss function. Recall that a decision rule for the 0-1 loss function will minimize the probability of error.

Solution: Recall that assuming a 0-1 loss function results in choosing the class to minimize the probability of error, which means choosing according to this rule:

$$\text{If } \frac{p(w_1|x)}{p(w_2|x)} > 1, \text{ choose 1}$$

Because there is a linear decision boundary, we search for the value such that we classify everything to the right as seabass, and everything to the left as salmon. This boundary is the value of x such that $p(w_1|x) = p(w_2|x)$.

$$p(w_1|x) = p(w_2|x) \implies 5p(x|w_1) = 7p(x|w_2)$$

$$\frac{5}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2}\frac{(x-5)^2}{\sigma^2}\right) = \frac{7}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2}\frac{(x-7)^2}{\sigma^2}\right)$$

$$\ln(5) - \frac{1}{2\sigma^2}(x-5)^2 = \ln(7) - \frac{1}{2\sigma^2}(x-7)^2$$

$$4\ln\left(\frac{5}{7}\right) - x^2 + 10x - 25 = -x^2 + 14x - 49$$

$$4\ln\left(\frac{5}{7}\right) + 24 = 4x$$

$$x = \ln\left(\frac{5}{7}\right) + 6 \approx 5.66$$

The decision rule is: If $x > 5.66$, classify as Seabass! Otherwise classify as Salmon.

Note: Because we had the same variance for both class conditionals, the x^2 term canceled out. If that was not the case, then there would be 3 regions, and we would allocate 2 of them to one fish, 1 of them to the other, depending on the height of the posterior probabilities. A good exercise would be to try to draw this: two 1-D Gaussians with different variances.

c) Now, find the decision rule using the loss matrix above. Recall that a decision rule, in general, minimizes the risk, or expected loss.

Solution: In the general case, we want to make the decision that minimizes risk. Thus, the decision boundary is located at where the risk of making either decision is equal, or:

$$R(\alpha_1|x) = R(\alpha_2|x)$$

Recall that $R(\alpha_i|x) = \sum_{j=1}^{C} \lambda_{ij} P(w = j|x)$.

$$\lambda_{11}P(w = 1|x) + \lambda_{12}P(w = 2|x) = \lambda_{21}P(w = 1|x) + \lambda_{22}P(w = 2|x)$$

$$2 * P(w = 2|x) = 1 * P(w = 1|x)$$

$$2 * \frac{7}{12}\mathcal{N}(7, 2) = 1 * \frac{7}{12}\mathcal{N}(5, 2)$$

Solving this like part b), we get that $x = 6 + \ln(\frac{5}{14}) \approx 4.97$. Thus, if the weight is greater than 4.97, we classify it as seabass and if not, we classify it as salmon.