Discussion 3 **Alex Francis**
CS 189 Email: afrancis@berkeley.edu
02/07/2017 OH: Wednesday 10:00am - 12:00pm (283E Soda)

From last time. Ask class if they're interested in going over the final question on the last worksheet. Write it on the board before class begins. The question, and general approach to matrix calculus is rather useful for Homework 2.

1 Matrix Calculus

Let $X \in \mathbb{R}^{n \times d}$ be a data matrix and $y \in \mathbb{R}^n$ be the corresponding vector of labels. What is the weight vector $\theta \in \mathbb{R}^d$ that minimizes the quadratic loss between the predicted labels $X\theta$ and the actual labels y?

Solution. Informally, we're asking for the hyperplane that minimizes the sum of squared distance from the hyperplane, or sum of squared errors. Intuitively, we know that the solution will therefore be the least squares estimate for the weight vector

$$\hat{\theta} = (X^\top X)^{-1} X^\top y$$

In order to more formally derive the result we formulate a convex optimization problem and utilize the techniques of matrix calculus. In this case, the problem can be stated precisely as,

$$\min J(\theta) = \|X\theta - y\|_2^2$$

Recall that $\|x\|_2^2 = x^\top x$. Therefore, we would like to minimize,

$$\begin{aligned}
(X\theta - y)^\top (X\theta - y) &= (\theta^\top X^\top - y^\top) && (1)\\
&= \theta^\top X^\top X\theta - y^\top X\theta - \theta^\top X^\top y + y^\top y && (2)\\
&= \theta^\top - 2\theta^\top X^\top y + y^\top y && (3)
\end{aligned}$$

Where (2) to (3) occurs because $y^\top X\theta = (X\theta)^\top y$. Then, we take the gradient and set equal to zero.

$$\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \left(\theta^\top X^\top X\theta - 2\theta^\top X^\top y + y^\top y\right)\\
&= \nabla_\theta \left(\theta^\top X^\top X\theta\right) - 2\nabla_\theta \left(\theta^\top X^\top y\right) + \nabla_\theta \left(y^\top y\right)
\end{aligned}$$

Pause derivation. The gradient of the sum is the sum of the gradient. We will tackle each of these terms individually, from right to left (in order of difficulty).

Easiest.

$$\nabla_\theta \left(y^\top y\right) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \sum_{i=1}^n y_i^2 \\ \vdots \\ \frac{\partial}{\partial \theta_d} \sum_{i=1}^n y_i^2 \end{bmatrix} = \vec{0}$$

Harder (Hint: Homework 2, 2(a)). Let's simplify the problem. $X^\top y = v$ for some vector $v \in \mathbb{R}^n$. Then, we have,

$$\begin{aligned}
\nabla_\theta \left(\theta^\top X^\top y\right) &= \nabla_\theta \left(\theta^\top v\right) \\
&= \begin{bmatrix} \frac{\partial}{\partial \theta_1} \sum_{i=1}^n \theta_i v_i \\ \vdots \\ \frac{\partial}{\partial \theta_d} \sum_{i=1}^n \theta_i v_i \end{bmatrix} \\
&= \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} \\
&= v
\end{aligned}$$

So, $\nabla_\theta \left(\theta^\top X^\top y\right) = X^\top y$

Hardest (Hint: Homework 2, 2(b)). It can be verified that (do this one on your own!),

$$\nabla_\theta \left(\theta^\top X^\top X \theta\right) = 2X^\top X \theta$$

General Techniques for Matrix Calculus.

1. Use expanded (summation) formulas and definition of gradient (over matrices or vectors). Then try to find patterns in resulting expression.
2. Use Fréchet derivative method for computing gradients. [1]

Using the above results, we have,

$$\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \left(\theta^\top X^\top X \theta - 2\theta^\top X^\top y + y^\top y\right) \\
&= 2X^\top X \theta - 2X^\top y
\end{aligned}$$

Setting this expression equal to zero, we have,

$$\hat{\theta} = (X^\top X)^{-1} X^\top y$$

2 Lecture on Decision Theory

I will not beat the point of the examples set in class further. For geometric intuition and definitions of loss functions and risk, see the lecture. We'll look at this at the mile-high level today.

The idea behind decision theory is to utilize some important techniques in probability to estimate the probability that a data point or feature vector is classified in a certain way given it's form/value. As a way of getting started, the theory here depends greatly on a single significant

[1] See Piazza: `https://piazza.com/class/ixwkta02en780?cid=233`

result - Bayes' Theorem.

$$\begin{aligned}
\text{posterior} &\propto \text{likelihood} \times \text{prior} \\
\mathbb{P}\{Y \mid X\} &= \frac{\mathbb{P}\{X \mid Y\}\mathbb{P}\{Y\}}{\mathbb{P}\{X\}} \\
&= \frac{\mathbb{P}\{X \mid Y\}\mathbb{P}\{Y\}}{\mathbb{P}\{X \mid Y\}\mathbb{P}\{Y\} + \mathbb{P}\{X \mid Y^C\}\mathbb{P}\{Y^C\}}
\end{aligned}$$

In the case of binary classification, i.e. when $Y = i \in \{0, 1\}$, we can solve easily for the form of the posterior on some class (let's use class 1 for now).

$$\mathbb{P}\{Y = 1 \mid X = x\} = \frac{\mathbb{P}\{X = x \mid Y = 1\}\mathbb{P}\{Y = 1\}}{\mathbb{P}\{X = x \mid Y = 1\}\mathbb{P}\{Y = 1\} + \mathbb{P}\{X = x \mid Y = 0\}\mathbb{P}\{Y = 0\}} \tag{1}$$

$$= \left(\frac{\mathbb{P}\{X = x \mid Y = 1\}\mathbb{P}\{Y = 1\}}{\mathbb{P}\{X = x \mid Y = 1\}\mathbb{P}\{Y = 1\}}\right)\left(\frac{1}{1 + \dfrac{\mathbb{P}\{X = x \mid Y = 0\}\mathbb{P}\{Y = 0\}}{\mathbb{P}\{X = x \mid Y = 1\}\mathbb{P}\{Y = 1\}}}\right) \tag{2}$$

$$= \frac{1}{1 + \dfrac{\mathbb{P}\{X = x \mid Y = 0\}\mathbb{P}\{Y = 0\}}{\mathbb{P}\{X = x \mid Y = 1\}\mathbb{P}\{Y = 1\}}} \tag{3}$$

Recall from high school (Hares and Wolves, anyone?) that the logistic function takes the form,

$$f(x) = \frac{1}{1 + e^{-g(x)}}$$

It can be shown that probability densities of the form,

$$p(x) = \alpha e^{-g(x)}$$

For some constant α and some function $g(x)$ yields the logistic equation for the posterior above.

3 Logistic Posterior with Different Variances

We have seen in class that Gaussian class conditionals with the same variance lead to a logistic posterior. Now we will consider the case when the class conditionals are Gaussian, but have different variance, i.e.

$$\begin{aligned}
X \mid Y = i &\sim \mathcal{N}(\mu_i, \sigma_i^2) \text{ where } i \in \{0, 1\} \\
Y &\sim \text{Bernoulli}(\pi)
\end{aligned}$$

Show that the posterior distribution of the class label given X is also a logistic function, however with a quadratic argument in X. Assuming 0-1 loss, what will the decision boundary look like (i.e. describe what the posterior probability plot looks like)?

Solution. From (3), we have, plugging in...

$$\begin{aligned}
\mathbb{P}\{Y=1 \mid X=x\} &= \frac{1}{1+\dfrac{\mathbb{P}\{X=x \mid Y=0\}\mathbb{P}\{Y=0\}}{\mathbb{P}\{X=x \mid Y=1\}\mathbb{P}\{Y=1\}}} \\
&= \frac{1}{1+\left(\dfrac{\dfrac{1}{\sigma_0\sqrt{2\pi}}\exp\left(-\dfrac{1}{2}\left(\dfrac{x-\mu_0}{\sigma_0}\right)^2\right)(1-\pi)}{\dfrac{1}{\sigma_1\sqrt{2\pi}}\exp\left(-\dfrac{1}{2}\left(\dfrac{x-\mu_1}{\sigma_1}\right)^2\right)(\pi)}\right)} \\
&= \frac{1}{1+\dfrac{\sigma_0(1-\pi)}{\sigma_1\pi}\left(\dfrac{\exp\left(-\dfrac{1}{2}\left(\dfrac{x-\mu_0}{\sigma_0}\right)^2\right)}{\exp\left(-\dfrac{1}{2}\left(\dfrac{x-\mu_1}{\sigma_1}\right)^2\right)}\right)} \\
&= \frac{1}{1+\dfrac{\sigma_0(1-\pi)}{\sigma_1\pi}\left(\exp\left(\left(\dfrac{1}{2\sigma_1^2}-\dfrac{1}{2\sigma_0^2}\right)x^2+\left(\dfrac{\mu_0}{\sigma_0^2}-\dfrac{\mu_1}{\sigma_1^2}\right)x+\left(\dfrac{\mu_1^2}{2\sigma_1^2}-\dfrac{\mu_0^2}{2\sigma_0^2}\right)\right)\right)}
\end{aligned}$$

This is of the form,

$$\frac{1}{1+\dfrac{\sigma_0(1-\pi)}{\sigma_1\pi}e^{-q_1(x)}}$$

Nothing that we can take place the logarithm of the constant in front of the e in the sum $q_1(x)$ (and that this does not change the degree of that polynomial) gives the result. In order to verify the result on the posterior for $Y=0$, let $q_0(x) = -q_1(x)$, and use the equality

$$\mathbb{P}\{Y=0 \mid X\} = 1 - \mathbb{P}\{Y=1 \mid X\}$$

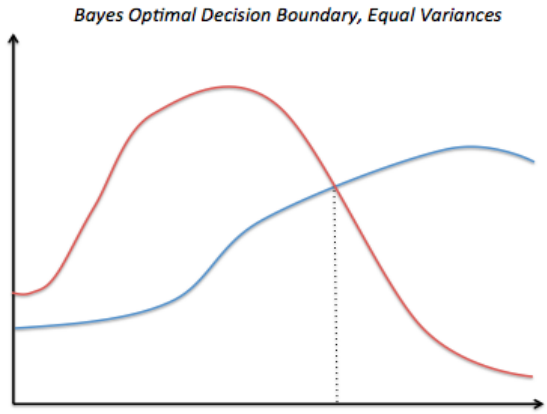
I defer to lecture on the fact that since we assume zero-one loss, we have that the form of the decision boundary is,

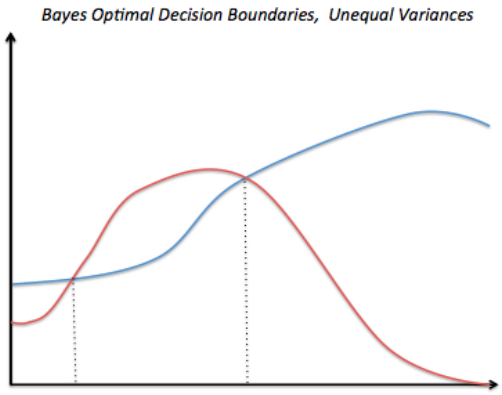
$$\mathbb{P}\{Y=1 \mid X\} = \mathbb{P}\{Y=0 \mid X\} = 1/2$$

Therefore,

$$\begin{aligned}
\frac{1}{1+e^{-q(x)}} &= \frac{1}{2} \\
e^{-q(x)} &= 1 \\
q(x) &= 0
\end{aligned}$$

We can solve for the roots of a quadratic and obtain one or two distinct roots. If there is one distinct root, the form of the boundary is similar to that discussed in class. Otherwise, the form of the boundary is as follows: there are two places at which the decision changes; therefore, the classifier must change it's decision from one class to the other, then back to the original class. See the visualization below.





4 Logistic Posterior with Exponential Class Conditionals

See the staff solution. I doubt we will get to this in section, as it is computationally intensive and similar to the previous problem.

5 Bayesian Decision Theory: Case Study

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 ($Y = 1$) and seabass ($Y = 2$). 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.:

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

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:

Table 1: *

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

1. First, compute the sample mean $\hat{\mu}_i$ and variance $\hat{\sigma}_i^2$ for the univariate Gaussian in both the seabass and the salmon case. Also compute the empirical estimates of the priors $\hat{\pi}_i$.

Solution.

$$\hat{\mu_1} = 1/5(3+4+5+6+7) = 5$$
$$\hat{\mu_2} = 1/7(5+6+7+8+9+7+\sqrt{2}+7-\sqrt{2}) = 7$$
$$\hat{\sigma_1}^2 = 1/5(2^2+1^2+0^2+1^2+2^2) = 2$$
$$\hat{\pi_1} = 5/12$$
$$\hat{\pi_2} = 7/12$$

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

   Solution. They're equal, meaning the decision boundary will be linear! (See 2 in this document, 1 on the original worksheet).

3. 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. If the decision rule relies on 0-1 loss, we need to find the point at which the posterior probabilities are equal (read: the x value). We can therefore set the posteriors equal to each other, or we can set one of them equal to $1/2$ and solve for x. I chose the first approach.

$$\mathbb{P}\{Y=1 \mid X\} = \mathbb{P}\{Y=2 \mid X\}$$
$$\mathbb{P}\{X=x \mid Y=1\}\mathbb{P}\{Y=1\} = \mathbb{P}\{X=x \mid Y=2\}\mathbb{P}\{Y=2\}$$
$$\mathbb{P}\{X=x \mid Y=1\}\hat{\pi_1} = \mathbb{P}\{X=x \mid Y=2\}\hat{\pi_2}$$
$$5\mathbb{P}\{X=x \mid Y=1\} = 7\mathbb{P}\{X=x \mid Y=2\}$$
$$5\left(\frac{1}{2\sqrt{2\hat{\sigma_1}^2}}\exp\left(-\frac{1}{2}\left(\frac{x-\hat{\mu_1}}{\hat{\sigma_1}}\right)^2\right)\right) = 7\left(\frac{1}{2\sqrt{2\hat{\sigma_2}^2}}\exp\left(-\frac{1}{2}\left(\frac{x-\hat{\mu_2}}{\hat{\sigma_2}}\right)^2\right)\right)$$
$$\exp\left(-\frac{1}{2}\left(\frac{x-5}{2}\right)^2\right) = \frac{7}{5} - \exp\left(-\frac{1}{2}\left(\frac{x-7}{2}\right)^2\right)$$
$$-\frac{1}{8}(x-5)^2 = \log\left(\frac{7}{5}\right) - \frac{1}{8}(x-7)^2$$
$$x^2 - 10x + 25 = x^2 - 14x + 49 - 8\log\left(\frac{7}{5}\right)$$
$$4x - 24 = -8\log\left(\frac{7}{5}\right)$$
$$x = 6 - 2\log\left(\frac{7}{5}\right)$$
$$x \approx 5.66$$

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

This is a more general version of the problem we just solved. Recall from lecture that the risk of a decision rule r is defined as,

$$\begin{aligned} R(r) &= \mathbb{E}[L(r(X), Y)] \\ &= P(Y = 1) \int L(r(x), 1) P(X = x \mid Y = 1) dx \\ &+ P(Y = 2) \int L(r(x), 2) P(X = x \mid Y = 2) dx \end{aligned}$$

The decision boundary occurs where,

$$L(r(x), 1) P(X = x \mid Y = 1) = L(r(x), 2) P(X = x \mid Y = 2)$$

Therefore, we have,

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

Which we can solve using the methods in the previous part to obtain $x = 6 + \ln(\frac{5}{14}) \approx 4.97$.

6 Circular Distributions

Unfortunately, it is highly unlikely that we will make it to this problem. See the staff solution!