10-601B Recitation 2

Calvin McCarter

September 10, 2015

1 Least squares problem

In this problem we illustrate how gradients can be used to solve the least squares problem.

Suppose we have input data matrix $X \in \mathbb{R}^{n \times p}$, output data $y \in \mathbb{R}^n$ and weight vector $w \in \mathbb{R}^p$, where p is the number of features per observation. The linear system $Xw = y$ corresponds to choosing a weight vector w that perfectly predicts y_i given $X_{\{}i, :\}$ for all observations $i = 1, \ldots, n$. The least squares problem arises out of the setting where the linear system $Xw = y$ is overdetermined, and therefore has no solution. This frequently occurs when the number of observations is greater than the number of features. This means that the outputs in y cannot be written exactly in terms of the inputs X. So instead we do the best we can by solving the least squares problem:

$$\min_w \|Xw - y\|_2^2.$$

We first re-write the problem:

$$
\begin{aligned}
&\min_w \|Xw - y\|_2^2 \\
&\min_w (Xw - y)^T (Xw - y) \\
&\min_w w^T X^T X w - w^T X^T y - y^T X w + y^T y \\
&\min_w w^T X^T X w - y^T X w - y^T X w + y^T y \qquad \text{using } a = a^T \text{ if } a \text{ is scalar, since } w^T X^T y \text{ is scalar} \\
&\min_w w^T X^T X w - 2y^T X w + y^T y
\end{aligned}
$$

To find the minimum, we find the gradient and set it to zero. (Recall that $\|Xw - y\|_2^2$ maps a p-dimensional vector to a scalar, so we can take its gradient, and the gradient is p-dimensional.) We apply the rules $\nabla_x \Big[x^T A x \Big] = 2Ax$ (where A is symmetric) and $\nabla_x \Big[c^T x \Big] = c$ proven in last recitation:

$$
\begin{aligned}
\nabla_w \Big[w^T X^T X w - 2y^T X w + y^T y \Big] =& \vec{0} \\
2X^T X w - 2X^T y =& \vec{0} \\
X^T X w =& X^T y.
\end{aligned}
$$

Recall that $X^T X$ is just a matrix, and $X^T y$ is just a vector, so w once again is the solution to a linear system. But unlike $Xw = y$, which had n equations and p unknowns, here we have p equations and p unknowns, so there will be at least one solution. In the case where $X^T X$ is invertible, we have $w = (X^T X)^{-1} X^T y = X^{-1}(X^T)^{-1} X^T y = X^{-1} y$, so we recover the solution to $Xw = y$. Otherwise, we can choose any one of the infinite number of solutions, for example $w = (X^T X)^+ Xy$, where A^+ denotes the pseudoinverse of A.

2 Matlab tutorial

If you missed recitation and aren't familiar with Matlab, please watch the first 27 minutes of this video: 10-601 Spring 2015 Recitation 2.

Here are the commands I used

```
3+4
x = 3
x = 3;
y = 'hello';
y = sprintf('hello world \%i \%f', 1, 1.5);
disp(y)

zeros(3,4)
eye(3)
ones(5)
rand(2,3)
A = 1+2*rand(2,3)
randn(4,1)
mu = 2; stddev = 3; mu + stddev*randn(4,1)
size(A)
numel(A)
who
whos
clear

A = rand(10,5)
A(2,4)
A(1:5,:)
subA = A([1 2 5], [2 4])
A(:,1) = zeros(10,1);
size(A(:))

X = ones(5,5);
Y = eye(5);
X'
inv(X)
X * Y
```

```
X .* Y
log(A)
abs(A)
max(X, Y)
X.^2
sum(A)
sum(A,1)
sum(A,2)
max(A,[],1)
max(A,[],2)

v = rand(5,2)
v>0.5
v(v>0.5)
index=find(v>0.5)
v(index)
[row_ix, col_ix] = find(v>0.5)
v(row_ix,col_ix)

for i=1:10
    disp(i)
end
x = 5;
if (x < 10)
    disp('hello');
elseif (x>10)
    disp('world');
else
    disp('moon');
end
clear
load('ecoli.mat');
imagesc(xTrain);
plot(xTrain(:,1));
```

3 MAP estimate for the Bernoulli distribution

3.1 Background

The probability distribution of a Bernoulli random variable X_i parameterized by μ is:

$$p(X_i = 1; \mu) = \mu \text{ and } p(X_i = 0; \mu) = 1 - \mu$$

We can write this more compactly (verify for yourself!):

$$p(X_i; \mu) = \mu^{X_i}(1-\mu)^{1-X_i}, \quad X_i \in 0, 1.$$

Also, recall from lecture that for a dataset with n iid samples, we have:

$$p(\mathbf{X}; \mu) = p(X_1, \ldots, X_n; \mu) = \prod_{i=1}^{n} p(X_i; \mu) = \mu^{\sum X_i}(1-\mu)^{\sum(1-X_i)}$$

$$\log p(\mathbf{X}; \mu) = \sum_{i=1}^{n} \Big[X_i \log \mu + (1 - X_i) \log(1-\mu) \Big]. \tag{1}$$

Finally, recall that we found the MLE by taking the derivative and setting to 0:

$$\frac{\partial}{\partial \mu} \log p(\mathbf{X}; \mu) = \frac{1}{\mu} \sum X_i - \frac{1}{1-\mu} \sum (1 - X_i) = 0 \tag{2}$$

$$\Rightarrow \hat{\mu}_{MLE} = \frac{\sum X_i}{n} = \frac{\#\text{ of heads}}{\#\text{ of flips}}$$

3.2 MAP estimation

In the previous section μ was an unknown but fixed parameter. Now we consider μ a random variable, with a prior distribution $p(\mu)$ and a posterior distribution after observing the coin flips $p(\mu|\mathbf{X})$. We're going to find the peak of the posterior distribution:

$$\begin{aligned} \hat{\mu}_{MAP} &= \operatorname*{argmax}_{\mu} p(\mu|\mathbf{X}) \\ &= \operatorname*{argmax}_{\mu} \frac{p(\mathbf{X}|\mu)p(\mu)}{p(\mathbf{X})} \\ &= \operatorname*{argmax}_{\mu} p(\mathbf{X}|\mu)p(\mu) \\ &= \operatorname*{argmax}_{\mu} \log p(\mathbf{X}|\mu) + \log p(\mu) \end{aligned}$$

So now we find the MAP estimate by taking the derivative and setting to 0:

$$\frac{\partial}{\partial \mu} \Big[\log p(\mathbf{X}; \mu) + \log p(\mu) \Big] = 0$$

Because for $\log p(\mathbf{X}|\mu)$ we use Eq. (1) above, we'll be able to use Eq. (2) for $\frac{\partial}{\partial \mu} \log p(\mathbf{X}|\mu)$.

For $\log p(\mu)$ we first need to specify our prior. We use the Beta distribution:

$$p(\mu) = \frac{1}{B(\alpha, \beta)} \mu^{\alpha-1}(1-\mu)^{\beta-1}$$

$$\log p(\mu) = \frac{1}{B(\alpha, \beta)} + (\alpha - 1)\log(\mu) + (\beta - 1)\log(1-\mu)$$

where $B(\alpha, \beta)$ is a nasty function that does not depend on μ. (It just normalizes $p(\mu)$ so that the total probability is 1.) Now we can find $\frac{\partial}{\partial \mu} \log p(\mu)$:

$$
\begin{aligned}
&\frac{\partial}{\partial \mu}\Big[\frac{1}{B(\alpha, \beta)} + (\alpha - 1)\log(\mu) + (\beta - 1)\log(1-\mu)\Big] \\
=&0 + (\alpha - 1)\frac{1}{\mu} + (\beta - 1)\frac{1}{1-\mu}(-1) \\
=&\frac{1}{\mu}(\alpha - 1) - \frac{1}{1-\mu}(\beta - 1).
\end{aligned}
$$

Finally, we compute our MAP estimate:

$$
\begin{aligned}
\Big[\frac{1}{\mu}\sum X_i - \frac{1}{1-\mu}\sum(1 - X_i)\Big] + \Big[\frac{1}{\mu}(\alpha - 1) - \frac{1}{1-\mu}(\beta - 1)\Big] &= 0 \\
\frac{1}{\mu}\Big(\sum(X_i) + \alpha - 1\Big) - \frac{1}{1-\mu}\Big(\sum(1 - X_i) + \beta - 1\Big) &= 0 \\
\Rightarrow \hat{\mu}_{MAP} = \frac{\sum X_i + \alpha - 1}{n + \beta + \alpha - 2} &= \frac{\text{\# of heads} + \alpha - 1}{\text{\# of flips} + \beta + \alpha - 2}
\end{aligned}
$$

3.3 Interpreting the Bayesian estimator

One way of interpreting the MAP estimate is that we pretend we had $\beta + \alpha - 2$ extra flips, out of which $\alpha - 1$ came up heads and $\beta - 1$ came up tails.

If $\alpha = \beta = 1$, $\hat{\mu}_{MAP} = \hat{\mu}_{MLE}$. In cases like this where our prior leads us to recover the MLE, we call our prior "uninformative". It turns out that Beta($\alpha = 1, \beta = 1$) reduces to a uniform distribution over $[0, 1]$, which lines up with our intuition about what an unbiased prior would look like!

Now suppose $\alpha = \beta = 10$, and we flip 3 heads out of 4 flips. We have $\hat{\mu}_{MLE} = 0.75$, but $\hat{\mu}_{MAP} = \frac{3+9}{4+18} \approx 0.55$. This prior corresponds to a belief that the coin is fair.

Now suppose $\alpha = \beta = 0.5$, and we flip 3 heads out of 4 flips. We have $\hat{\mu}_{MLE} = 0.75$, but $\hat{\mu}_{MAP} = \frac{3-0.5}{4-0.5} \approx 0.83$. Our prior is pulling our estimate *away* from $\frac{1}{2}$! This prior corresponds to a belief that the coin is unfair (maybe it's a magician's coin) but we have no idea which way it's bent.

For a fixed α, β prior, what happens as we get more samples?

$$
\begin{aligned}
&\lim_{n\to\infty} \hat{\mu}_{MAP} \\
=&\lim_{n\to\infty} \frac{n\mu + \alpha - 1}{n + \beta + \alpha - 2} \\
=&\mu
\end{aligned}
$$

In other words, the MAP estimate converges like the MLE estimate to the true μ, and the effect of our prior diminishes.