# 1 Multinomial MLE

Suppose we have multiple classes $a_1, ..., a_n$ and m data samples $\mathcal{D}$. We seek to compute $p(\theta|D)$ where $\theta = (p_1, ..., p_n)$ with $\sum p_i = 1$. Bayes rule and MLE tell us to find the optimal parameters $\theta$ maximizing $p(D|\theta)$.

Write the expression for MLE we are seeking to maximize

$$log(\prod_{i=1}^{m} p(D^i|\theta)) = log(\prod_{j=1}^{n} K_j p_j^{\alpha_j} =$$

$$\sum_{j=1}^{n} log(K_j) + \sum_{j=1}^{n} \alpha_j log(p_j) = f(p)$$

where we product over the classes and $K_j$ corresponds to a binomial coefficient

How do we go about maximizing this expression? Compute the maximizer.

It is most convenient to use the method of lagrange multipliers, since we must also enforce the constraint $g(p) = \sum_{j=1}^{n} p_j = 1$. Thus we are solving

$$\nabla f(p) - \lambda \nabla g(p) = 0 \implies$$

coordinate wise

$$\frac{\alpha_j}{p_j} - \lambda \implies \forall i, j, \frac{\alpha_j}{p_j} = \frac{\alpha_i}{p_i}$$

Then we solve for $p_1$(and in general $p_i$):

$$p_j = \frac{p_1}{x_1} x_i \implies \sum p_j = 1 \implies p_1 + \sum_{i \neq 1} \frac{p_1}{x_1} x_i = 1 \implies$$

$$p_1 = \frac{x_1}{\sum x_i}$$

as we would intuitively expect

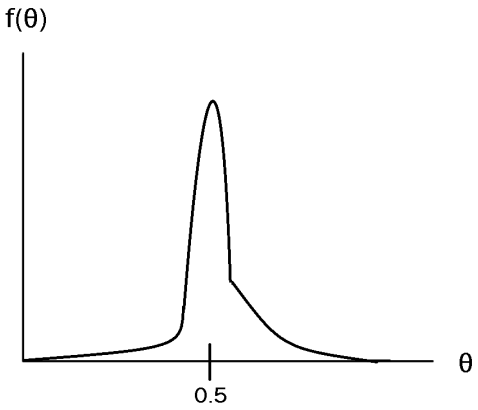
What do we know about the maximizer?

The expression is concave because it is the sum of concave functions.

# 2 MAP Bernoulli

Now let's consider MAP instead. Recall that we're trying to maximize the posterior probability, which is proportional to likelihood * prior. In mathematical form, $p(\theta|\mathcal{D}) \propto \prod p(\mathcal{D}^{(n)}|\theta)p(\theta)$.

The prior probability distribution given in class is as follows:




where f($\theta$=0.25) = 0.1, f($\theta$=0.5)=0.6, and f($\theta$=0.75) = 0.2.

### 2.1 MAP estimate with 2 samples

Let us consider only the first 2 samples, which we learn are both heads. Formally solving for $\theta$ is similar to the process for MLE, except the prior probability is also introduced. For simplicity of this exercise, just find and compare the posterior probabilities corresponding to $\theta = 0.25$, 0.5 and 0.75. Which $\theta$ gives you the highest probability?

$p(\theta = 0.25|\mathcal{D}) = 0.25^2 * 0.1 = 0.0063$
$p(\theta = 0.5|\mathcal{D}) = 0.5^2 * 0.6 = 0.15$
$p(\theta = 0.75|\mathcal{D}) = 0.75^2 * (0.05) = 0.11$

$\theta = 0.5$ gives you the highest probability.

### 2.2 MAP estimate with 10 samples

Now consider all 10 samples, and find the corresponding posterior probabilities for the three $\theta$ values again. Which $\theta$ gives you the highest probability?

$p(\theta = 0.25|\mathcal{D}) = 0.25^8(1 - 0.25)^2 * 0.1 = 8.58 * 10^{-7}$
$p(\theta = 0.5|\mathcal{D}) = 0.5^8(1 - 0.5)^2 * 0.6 = 0.00059$
$p(\theta = 0.75|\mathcal{D}) = 0.75^8(1 - 0.75)^2 * 0.2 = 0.0013$

$\theta = 0.75$ gives you the highest probability.

### 2.3 Effect of number of samples on MAP estimate

How do the $\theta$ values from parts $a$ and $b$ compare? In which case does prior probability play a bigger role, and why?

The $\theta$ value changed. As more samples are introduced, the prior probability becomes less impactful. More weight is instead put on the observed data.

# 3 MLE and MAP with Other Distributions

In general given a distribution $d$ on $p(D|\theta)$ how do we go about computing the posterior distribution $p(\theta|D)$ with a prior on $p(\theta)$?

Perform MLE on the expression $p(D|\theta)p(\theta)$ instead of just MLE on $p(D|\theta)$. MLE will result in optimal parameters for $\theta$(given the data we have) which will allow us to compute $p(\theta|D)$ for classification.

MLE and MAP for Gaussian Distribution:
We have the MLE Formulation

$$\hat{\mu}\frac{1}{n}\sum_{j=1}^{n}x_j$$

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum(x_j - \hat{\mu}_{MLE})$$

Why is the estimate for the variance biased?

Because its estimate is based on the estimate of the MLE mean instead of using the true mean of the dataset.

We also have the MAP Formulation
Given a prior distribution on our mean $\mu$ $N(\eta, \lambda^2)$ we have the posterior MAP distribution on mean:

$$\hat{\mu}_{MAP} = \frac{\frac{1}{\sigma^2}\sum_{i=1}^{n}x_i + \frac{\eta}{\lambda^2}}{\frac{n}{\sigma^2} + \frac{1}{\lambda^2}}$$

# 4 MLE in Practice: Text Classification

Say we have a bunch of movie reviews we want to classify into positive or negative categories. But we don't want to spend the time writing and training a high powered neural network(or have the data for that matter). What's a nice alternative?

Naive Bayes!

Describe how we could do this using a simple bag of words feature extraction method.

The bag of words model causes allows us to use word frequency counts as features for a given text. We establish a dictionary of 10000 words and count how many times each one occurs in the text.

Once we have our features we can "train" an MLE Naive Bayes model by for each class, positive or negative, counting the number of instances of each word in our dictionary given a dataset D. As instructed by MLE to compute $p(D|\theta)$. Then assuming a bernoulli distribution on $p(\theta)$ where $\theta$ is the class, we simply count the total number of negative vs positive reviews.

To classify compute the probability $\prod P(\theta|x_i)P(\theta)$