Logistic Regression

In this lecture we will learn about the discriminative counterpart to the Gaussian Naive Bayes (for continuous features).

Machine learning algorithms can be (roughly) categorized into two categories:

- *Generative* algorithms, that estimate $P(\vec{x}_i, y)$ (often they model $P(\vec{x}_i|y)$ and $P(y)$ separately).
- *Discriminative* algorithms, that model $P(y|\vec{x}_i)$

The Naive Bayes algorithm is *generative*. It models $P(\vec{x}_i|y)$ and makes explicit assumptions on its distribution (e.g. multinomial, categorical, Gaussian, ...). The parameters of this distributions are estimated with MLE or MAP. We showed previously that for the Gaussian Naive Bayes $P(y|\vec{x}_i) = \frac{1}{1+e^{-y(\vec{w}^T\vec{x}+b)}}$ for $y \in \{+1, -1\}$ for specific vectors $\vec{w}$ and b that are uniquely determined through the particular choice of $P(\vec{x}_i|y)$.

Logistic Regression is often referred to as the *discriminative* counterpart of Naive Bayes. Here, we model $P(y|\vec{x}_i)$ and assume that it takes on exactly this form

$$P(y|\vec{x}_i) = \frac{1}{1 + e^{-y(\vec{w}^T\vec{x}+b)}}.$$

We make little assumptions on $P(\vec{x}_i|y)$, e.g. it could be Gaussian or Multinomial. Ultimately it doesn't matter, because we estimate the vector $\vec{w}$ and b directly with MLE or MAP.

For a lot more details, I strongly suggest that you read this excellent book chapter by Tom Mitchell

Maximum likelihood estimate (MLE)

In MLE we choose parameters that **maximize the conditional likelihood**. The conditional data likelihood $P(\vec{y} \mid X, \vec{w})$ is the probability of the observed values $\vec{y} \in \mathbb{R}^n$ in the training data conditioned on the feature values $\vec{x}_i$. Note that $X = [\vec{x}_1, \ldots, \vec{x}_i, \ldots, \vec{x}_n] \in \mathbb{R}^{d \times n}$. We choose the paramters that maximize this function and we assume that the y_i's are independent given the input features $\vec{x}_i$ and $\vec{w}$. So,

$$P(\vec{y} \mid X, \vec{w}) = \prod_{i=1}^{n} P(y_i \mid \vec{x}_i, \vec{w}).$$

Now,

$$log\left(\prod_{i=1}^{n} P(y_i|\vec{x}_i, \vec{w})\right) = -\sum_{i=1}^{n} \log(1 + e^{-y_i\vec{w}^T\vec{x}})$$

Note, that we absorbed the parameter b into $\vec{w}$ through an additional constant dimension.

$$\begin{aligned}\hat{\vec{w}}_{MLE} &= \underset{\vec{w}}{\text{argmax}} -\sum_{i=1}^{n} \log(1 + e^{-y_i\vec{w}^T\vec{x}}) \\ &= \underset{\vec{w}}{\text{argmin}} \sum_{i=1}^{n} \log(1 + e^{-y_i\vec{w}^T\vec{x}})\end{aligned}$$

We need to estimate the parameters $\vec{w}$. To find the values of the parameters at minimum, we can try to find solutions for $\nabla_{\vec{w}} \sum_{i=1}^{n} \log(1 + e^{-y_i \vec{w}^T \vec{x}}) = 0$. This equation has no closed form solution, so we will use on the *negative log likelihood* $\ell(\vec{w}) = \sum_{i=1}^{n} \log(1 + e^{-y_i \vec{w}^T \vec{x}})$.

Maximum a Posteriori (MAP) Estimate

In the MAP estimate we treat $\vec{w}$ as a random variable and can specify a prior belief distribution over it. We may use: $\vec{w} \sim \mathcal{N}(\vec{0}, \sigma^2 I)$. This is the Gaussian approximation for LR.

Our goal in MAP is to find the *most likely* model parameters *given the data*, i.e., the parameters that **maximaize the posterior**.

$$P(\vec{w} \mid D) = P(\vec{w} \mid X, \vec{y}) \propto P(\vec{y} \mid X, \vec{w})\, P(\vec{w})$$

$$\hat{\vec{w}}_{MAP} = \underset{\vec{w}}{\operatorname{argmax}}\, log\ (P(\vec{y} \mid X, \vec{w}) P(\vec{w})) = \underset{\vec{w}}{\operatorname{argmin}} \sum_{i=1}^{n} \log(1 + e^{-y_i \vec{w}^T \vec{x}}) + \lambda \vec{w}^\top \vec{w}, '$$

where $\lambda = \frac{1}{2\sigma^2}$. Once again, this function has no closed form solution, but we can use Gradient Descent on the *negative log posterior* $\ell(\vec{w}) = \sum_{i=1}^{n} \log(1 + e^{-y_i \vec{w}^T \vec{x}}) + \lambda \vec{w}^\top \vec{w}$ to find the optimal parameters $\vec{w}$.

For a better understanding for the connection of Naive Bayes and Logistic Regression, you may take a peek at these excellent notes.