Machine Learning 1

Sample Questions 2 Solutions

WS18/19 (Sample questions for Friday, February 1st)

Note: Solutions can contain errors. Those will be fixed.

Q.1 (10 points)

Q.1.1.

1. Standard gradient descent has the iteration rule
$$\theta_{t+1} = \theta_t - \lambda \nabla f, \qquad \lambda \in \mathbb{R}$$
This iteratively finds the minimum of a function. It works, because the gradient indicates the region of steepest ascent. This can be seen if we use the definition of directional derivative and look which direction increases $f(\theta_{t+1})$ the most. Let $f: \mathbb{R}^m \to \mathbb{R}$ and $U \coloneqq \{\mathbf{u} \in \mathbb{R}^m : \|\mathbf{u} - \mathbf{x}\|^2 \leq \lambda\}$ and $\|\mathbf{v}\| = 1$. We then have:
$$\begin{aligned} \max_{\mathbf{u} \in U} f(\mathbf{u}) &\approx f(\mathbf{x}) + \lambda \langle \mathbf{v}, \nabla f(\mathbf{x}) \rangle, \qquad \lambda \in \mathbb{R} \\ &= f(\mathbf{x}) + |\mathbf{v}| \cdot |\nabla f(\mathbf{x})| \cdot \cos \alpha \\ &= f(\mathbf{x}) + 1 \cdot |\nabla f(\mathbf{x})| \cdot \cos \alpha \end{aligned}$$
Above is maximized for $\alpha = k \cdot 2\pi, with\ k \in \mathbb{Z}$. In other words, any direction v that maximizes $f(\mathbf{u})$ goes along the gradient. We minimize this function, if we set $\alpha = k \cdot \pi, with\ k \in \mathbb{Z}$, which means we travel in the opposite direction of the gradient. Therefore, we will continuously approach a local minimum of the function (if existent).
2. Momentum based gradient descent adds a velocity term to the update step. This makes the path a little bit more unpredictable, as it now not only depends on the current gradient but also previous geometry. In addition, we now have another global parameter controlling the velocity, growing our parameters to two.
3. Nesterov based gradient descent is similar to Momentum based gradient. In opposition to the latter, we evaluate the geometry at future steps in order to combat overshooting our minimum.

Q.1.2.

We know from the central limit theorem, that the sum of independent, identically distributed random variables converges to a normal distribution with vanishing variance. More generally, if we assume *exchangeability* of our data-sequence $X = X_1, X_2, \ldots, X_n$, we know from de'Finetti's exchangeability theorem, that X can be explained by a parameter θ with distribution $\mu(\theta)$. This distribution will decrease in variance the more samples we obtain. [Note: I have to recheck this explanation)].

Q.1.3.

Backpropagation assumed a function to be differentiable in its entire domain. Piecewise continuous functions such as Heaviside do not require that requirement for certain points. In addition, we want small changes in weights w_{ij} produce a small change in the output of the neural network (just like $f(x+\epsilon) \approx f(x) + \epsilon f'(x)$) in order to utilize gradient descent effectively. The Heaviside does not fulfill that requirement, as small changes either do not change the output or produce one sudden jump.

Q.1.4.

$$p(\theta|\mathbf{X}_{new}, \mathbf{X}_{old}) \propto p(\theta, \mathbf{X}_{old}, \mathbf{X}_{new}) \propto p(\theta|\mathbf{X}_{old}) p(\mathbf{X}_{old}|\theta)$$

Q.1.5.

Conjugate priors are prior distributions, that are in the same functional class as the Likelihood distributions. This allows us to easily specify the form of the posterior and is done out of pure convenience.

Improper priors are prior distributions that cannot be normalized (such as $Beta(0,0)$).

When selecting based on invariance principles, we pick a prior that satisfies certain invariance principles. Jeffreys prior for example, facilitates invariance with respect to reparameterization.

Q.2 (10 points)

Q.2.1.

As the Dirichlet distribution is conjugate to the categorial likelihood, we should not be surprise, that we will receive another Dirichlet distribution (or in this case, Beta) as posterior.

$$\begin{aligned} P(\theta|\mathbf{X}) &= \frac{P(\mathbf{X}|\theta)P(\theta)}{P(\mathbf{X})} \\ &\propto \underbrace{\theta^k(1-\theta)^{n-k}}_{propotional\ to\ likelihood} \cdot \underbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1}}_{beta\ prior} \\ &= \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1} \\ &= Beta(\theta; k+\alpha, n-k+\beta) \end{aligned}$$

Q.2.2.

If $1 = heads$, then:

Maximum Likelihood:

$$\theta_{MAX} = \frac{k}{n} = \frac{5}{8}$$

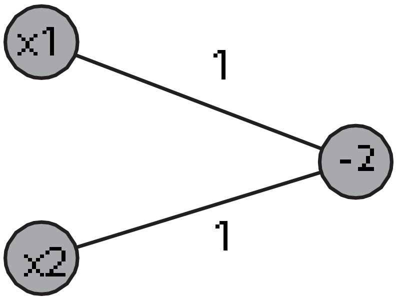
Bayesian Formulation:

Let $X \sim Beta(\theta; k+\alpha, n-k+\beta)$.
We have $k = 5, \alpha = 3, \beta = 7, n = 8$.
Therefore: $\alpha_{new} = 5 + 3 = 8, \beta_{new} = 8 - 5 + 7 = 10$

$$E[X] = \frac{\alpha_{new}}{\alpha_{new} + \beta_{new}} = \frac{8}{18}.$$

Q.2.3.

The following should suffice:



This is equal to:

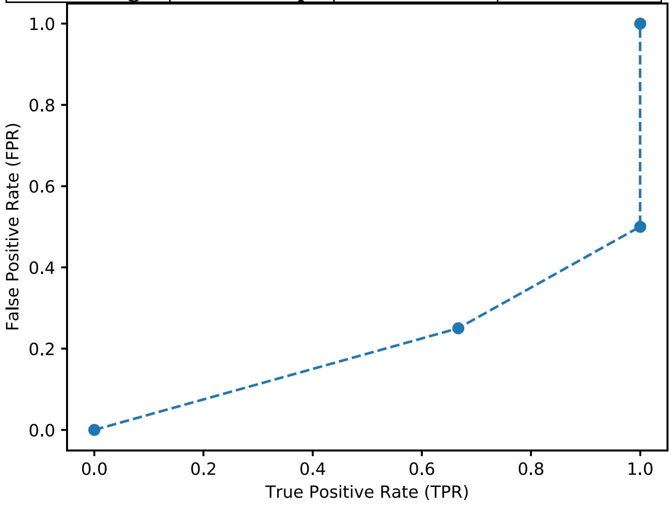
$$h_{out} = Heaviside(1 \cdot x_1 + 1 \cdot x_2 - 2)$$

Q.3

As our activation is sigmoid, we already know that the ROC curve contains the points $(0,0), (1,1)$ (for why, look up exercise 6.4 (I think), which shows, that $\sigma(x) = 1 \; or \; \sigma(x) = 0$ happens only in the limit). Now lets create some tables:

$\theta = 0.5$		True	
		C1	C2
Predicted	C1	3	2
	C2	0	2
$TPR = 1,$	$FPR = \frac{2}{4}$		

$\theta = 0.8$		True	
		C1	C2
Predicted	C1	2	1
	C2	1	3
$TPR = \frac{2}{3}$	$FPR = \frac{1}{4}$		


False Positive Rate (FPR)
True Positive Rate (TPR)